Freie Universität Berlin

Fachbereich Wirtschaftswissenschaft

Early Childhood and the Formation of Cognitive Skills

Three Empirical Essays in Education Economics

Inaugural Dissertation

vorgelegt von

Stephan Sievert, M.Sc.

zur Erlangung des akademischen Grades

doctor rerum politicarum

(Doktor der Wirtschaftswissenschaft)

Berlin, 2019

Gedruckt mit Genehmigung des Fachbereichs Wirtschaftswissenschaft
der Freien Universität Berlin

Dekan:	Prof. Dr. Dr. Andreas Löffler
Erstgutachterin:	Prof. Dr. C. Katharina Spieß *Freie Universität Berlin und DIW Berlin*
Zweitgutachter:	Prof. Dr. Jan Marcus *Universität Hamburg und DIW Berlin*
Tag der Disputation:	17. Juni 2020

ACKNOWLEDGMENTS

I want to thank all the people who have supported and helped me while writing this dissertation. In particular, I am indebted to my supervisor C. Katharina Spieß who "adopted" me as an external PhD student among her doctoral candidates at the Education and Family Department at DIW Berlin. I very much appreciate the extra work she put into guiding me through this dissertation over the years. Her valuable feedback and sometimes critical comments have considerably improved my work. Most importantly, however, I always felt that she believed in me completing this dissertation and gave much-needed encouragement.

I am also grateful to everyone else at the Education and Family Department that I had the chance to work with, discuss with, and in general spend time with. The somewhat coincidental decision two or three years into this project to regularly work at DIW Berlin (I wanted direct access to SOEP regional data instead of SOEP remote) instead of my home environment proved to be one of the best strategic course-settings. The informal learning that took place on these occasions was invaluable to me. Apart from this, I always enjoyed the atmosphere at the department. I felt welcome at our monthly doctoral colloquiums from day one. In particular, I would like to thank Maximilian Bach, my co-author of Chapter 4, as well as Felix Weinhardt, Frauke Peter, Jan Marcus, and Jan Bietenbeck (although not a DIW employee) who have read parts of this dissertation and provided valuable feedback. Jan Marcus gets another thank you for agreeing to be my second supervisor.

Berlin, October 2019

Stephan Sievert

CONTENTS

LIST OF TABLES

LIST OF FIGURES

RECHTLICHE ERKLÄRUNG

Erklärung gem. §4 Abs. 2 (Promotionsordnung)

Hiermit erkläre ich, dass ich mich noch keinem Promotionsverfahren unterzogen oder um Zulassung zu einem solchen beworben habe, und die Dissertation in der gleichen oder einer anderen Fassung bzw. Überarbeitung einer anderen Fakultät, einem Prüfungsausschuss oder einem Fachvertreter an einer anderen Hochschule nicht bereits zur Überprüfung vorgelegen hat.

Berlin, Oktober 2019
Stephan Sievert

Erklärung gem. §10 Abs. 3 (Promotionsordnung)

Hiermit erkläre ich, dass ich für die Dissertation folgende Hilfsmittel und Hilfen verwendet habe: Software LateX, Stata, Microsoft Excel, Microsoft Word, Literatur siehe Literaturverzeichnis. Auf dieser Grundlage habe ich die Arbeit selbstständig verfasst.

Berlin, Oktober 2019
Stephan Sievert

Ko-Autorenschaften und Vorveröffentlichungen

Chapter 2: The Effects of Universal Child Care Provision in Adolescence

- Ko-Autoren: keine
- Vorveröffentlichungen: keine

Chapter 3: Engaging Teaching Practices and Achievement – A Within-Student Approach in Three Subjects

- Ko-Autoren: keine
- Vorveröffentlichungen: keine

Chapter 4: Birth Cohort Size Variation and the Estimation of Class Size Effects

- Ko-Autoren: Maximilian Bach (ZEW Mannheim)
- Revise and Resubmit bei *Journal of Human Resources*
- Vorveröffentlichung: DIW Discussion Paper 1817 / 2019
- Teile dieses Kapitels sind erschienen in:
    - Bach, M. und Sievert, S. (2018): Kleinere Grundschulklassen können zu besseren Leistungen bei SchülerInnen führen, DIW Wochenbericht 22/2018.

ABSTRACT

This dissertation examines cognitive skill returns to different features of education systems in three independent research articles. It concentrates on interventions in early childhood, which is the age when children's cognition is particularly malleable. Each of the three articles makes an independent contribution to the economics of education literature.

Chapter 2 estimates the medium- and long-run effects of attending universal center-based child care. While short-run benefits of child care attendance especially for children from low socio-economic backgrounds are well-established in the literature, causal evidence on long-run outcomes is still patchy. The article fills a gap in the literature by focusing on a number of educational outcomes, most of which have not been causally studied before. These include secondary track choices, grade retentions, cognitive skill outcomes as well as aspirations towards further education.

The study draws on information from the German Socio-Economic Panel (SOEP), a large representative household survey that provides annual information on children's child care careers as well as rich personal background data. For identification, it exploits an amendment to the German "Child and Youth Welfare Act" (*Kinder- und Jugendhilfegesetz*) in 1992. The reform established a legal right to a heavily subsidized half-day place in child care for all children from the age of 3 until the beginning of primary school. The reform was meant to especially improve the situation for 3-year-old children who had been hit hardest by the prevalent situation of undersupply at the time, as places were often assigned by age. Exploiting the fact that the expansion in child care supply was staggered across counties for arguably exogenous reasons, the causal effect of one additional year of child care attendance is estimated in an instrumental variables framework where the level of regional child care supply serves as the excluded instrument.

The results indicate that German language grades in adolescence are positively influenced by longer child care attendance. The effect is particularly strong among weaker students as reflected by a sizably reduced likelihood of obtaining one of the three worst grades.

There is also evidence for increased educational aspirations, again most notably at the lower margin where students decide between pursuing a vocational degree upon completion of high school or not pursuing any further degree. Taken together, the results corroborate previous evidence that child care attendance is most beneficial for disadvantaged children. This strengthens the case for public funding of child care centers as a means to tackle inequalities in child development.

Chapter 3 focuses on the largely unanswered question what classroom actions by teachers are effective in conferring cognitive skills upon students. Specifically, the chapter assesses the effectiveness of primary school teachers' intentions to increase their students' engagement with the course content via different teaching practices. These practices include summarizing key messages, relating lessons to students' daily lives, use questioning, encouraging students, giving praise, and bringing interesting materials to class.

The analysis is based on a unique dataset that combines information from the 2011 waves of the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) for a representative sample of German fourth-graders. TIMSS is an achievement study that tests student competencies in mathematics and science, while PIRLS is dedicated to reading. The year 2011 marks the only occasion so far that the two studies have been sampled together. For this reason, test scores are observed in three different subjects for each student. This provides extra variation for the estimation of within-student between-subject models that form the empirical basis of this chapter. The main results indicate that engaging teaching practices have beneficial effects on students from low socio-economic backgrounds only. Averaged across subjects, it is estimated that a one-standard-deviation-increase on a composite scale measuring the use of potentially engaging teaching practices raises test scores by 4.6 percent of a standard deviation of the test score distribution. In subject-specific analyses based on correlated random effects models, the detected effect is largest in reading.

With regard to policy implications, it is important to note that the benefits for children from low socio-economic backgrounds are not offset by significantly lower achievement among other students. Therefore, it is concluded that greater use of engaging teaching practices in primary schools can – very much like longer child care attendance – serve as a vehicle to tackle inequalities in children's cognitive skill development.

Chapter 4 is dedicated to the estimation of class size effects in primary school. Class size is one of the key policy levers in education, as teachers' salaries account for the bulk of educational expenditures in many countries. As of yet, however, there is no academic consensus about the effects that class size reductions or increases generally have. In particular, effect sizes are typically much larger in experimental studies than in quasi-experimental setups. Chapter 4 is an attempt to reconcile these research findings.

The chapter shows how grade retentions of poorly performing students give rise to an upward bias in class size estimates that are based on within-school variation in cohort size over time. The bias, which depresses the typically negative class size effects toward zero, is produced by a negative mechanical relationship between initial cohort size and the share of previously retained students in the same cohort in higher grades. The existence of this compositional effect finds empirical support in administrative data on school enrollment for all primary schools in the German states of Saarland and Saxony. It is also shown that the resulting bias can be easily corrected by controlling for previous grade retentions. Performing this correction, class size effects are estimated in a dataset that covers four entire cohorts of students in Saarland who participated in state-wide centralized exams in language and mathematics at the end of grade 3. Instrumenting class size in grade 3 by predicted class size based on imputed cohort size, the results indicate that test scores are increased by around 1.9 and 1.4 percent of a standard deviation for each one-student decrease in class size. However, class size effects seem to be highly non-linear. Whereas language and math test scores are increased by 4.8 and 3.8 percent of a standard deviation in larger classes of more than 20.5 students for each student less, respectively, no effect is found in classes with fewer students. Significant heterogeneities are also observed with regard to student background. Again, disadvantaged students (those with insufficient German proficiency or a learning disability) benefit disproportionately strongly. Finally, evidence is found for a decreased likelihood of grade repetitions in smaller classes.

The results of Chapter 4 provide first causal evidence of significant class size effects on test scores in Germany. They suggest that class size reductions to increase student achievement should be targeted at larger classes. Conversely, class size in small classes may be increased up to a threshold of around 20.5 students per class with no adverse effects on achievement.

The three research articles of this dissertation are framed by Chapters 1 and 5. **Chapter 1** introduces the topic of cognitive skill formation and highlights the main contributions of this dissertation. **Chapter 5** concludes by critically discussing the main findings and outlining policy implications.

ZUSAMMENFASSUNG

Die vorliegende Dissertation untersucht die Wirkung unterschiedlicher Charakteristika von Bildungssystemen auf kognitive Fähigkeiten von Kindern und Jugendlichen. Sie konzentriert sich auf Maßnahmen, die vor dem 10. Lebensjahr eines Kindes ansetzen. Während dieser Phase lässt sich Kognition von Kindern besonders gut beeinflussen. Die drei Artikel dieser Dissertation leisten einen jeweils eigenständigen Beitrag zur Bildungsökonomie-Literatur.

Kapitel 2 untersucht mittel- und langfristige Effekte eines längeren Kindergartenbesuchs. Während positive kurzfristige Effekte insbesondere für Kinder mit niedrigem sozioökonomischen Status häufig nachgewiesen worden sind, ist die kausale Evidenz für Langfristeffekte noch lückenhaft. Die Studie schließt eine Lücke in der Literatur, indem sie sich auf eine Reihe von Bildungsergebnissen konzentriert, die bisher noch nicht mittels kausaler Analysen untersucht worden sind. Dazu gehören die Wahl der weiterführenden Schulart, Klassenwiederholungen, kognitive Fähigkeiten sowie weitere Bildungsaspirationen nach Abschluss der Sekundarschule.

Die Studie basiert auf Daten des Sozioökonomischen Panels (SOEP), einer repräsentativen Haushaltsbefragung, die jährliche Informationen über den Besuch von Kindertageseinrichtungen sowie umfangreiche persönliche Hintergrundmerkmale enthält. Zur Identifikation der Effekte nutzt sie eine Novelle des Kinder- und Jugendhilfegesetzes von 1992. Mit der Reform wurde ein rechtlicher Anspruch auf einen stark subventionierten Halbtagskindergartenplatz für alle Kinder von 3 Jahren bis zum Beginn der Grundschule eingeführt. Vor allem die Situation von dreijährigen Kindern sollte damit verbessert werden, da diese von der damals vorherrschenden Knappheit an Kindergartenplätzen am stärksten betroffen waren. Der kausale Effekt eines zusätzlichen Kindergarten-Jahres kann mittels eines Instrumentenvariablen-Ansatzes geschätzt werden, in dem das regionale Niveau des Betreuungsangebots als Instrument dient, da die Ausweitung des Betreuungsangebots aus exogenen Gründen regional unterschiedlich schnell erfolgte.

Die Ergebnisse zeigen, dass sich ein längerer Kindergartenbesuch positiv auf die Deutschnote im Jugendalter auswirkt. Unter schwächeren Schülern ist der Effekt besonders stark ausgeprägt, was sich an einer verringerten Wahrscheinlichkeit zeigt, eine der drei schlechtesten Noten zu erhalten. Außerdem gibt es Hinweise auf erhöhte Bildungsaspirationen schwächerer Schüler, die nach ihrem Abschluss vor der Wahl zwischen einer beruflichen Ausbildung und keiner weiteren Qualifizierung stehen. Zusammenfassend bestätigen die Ergebnisse bestehende Forschungsergebnisse darin, dass Kinderbetreuung am vorteilhaftesten für benachteiligte Kinder ist. Dies stärkt das Argument für eine öffentliche Finanzierung von Kindertageseinrichtungen als Mittel zur Bekämpfung von Ungleichheiten in der kindlichen Entwicklung.

Kapitel 3 konzentriert sich auf die weitgehend unbeantwortete Frage, welche Unterrichtsmethoden sich besonders dafür eignen, Schülern kognitive Fähigkeiten zu vermitteln. Insbesondere wird die Wirksamkeit unterschiedlicher Lehrmethoden in der Grundschule bewertet, die darauf abzielen, das Engagement von Schülern zu erhöhen, aktiv am Unterrichtsgeschehen teilzunehmen. Zu diesen Methoden gehört das Zusammenfassen von Kernbotschaften, das Herstellen von Bezügen zum täglichen Leben der Schüler, das gezielte Nachfragen, das Ermutigen, das Loben sowie das Mitbringen interessanter Unterrichtsmaterialien.

Die Analyse basiert auf einem Datensatz, der für eine repräsentative Stichprobe deutscher Viertklässler Informationen aus den 2011er Wellen der TIMSS- und IGLU-Studien kombiniert. TIMSS ist eine Schulleistungsuntersuchung, die die Kompetenzen von Schülern in Mathematik und Naturwissenschaften testet, während sich IGLU dem Lesen widmet. Im Jahr 2011 sind die beiden Studien zum bislang einzigen Mal gemeinsam durchgeführt worden. Daher lassen sich für jeden Schüler Testergebnisse in drei verschiedenen Fächern beobachten. Dies bietet zusätzliche Variation für den *within-student between-subjects*-Ansatz, der die empirische Grundlage dieses Kapitels bildet. Die Ergebnisse von Modellen mit fixen Schülereffekten zeigen, dass engagierende Lehrmethoden nur für Schüler mit niedrigem sozio-ökonomischen Status positive Auswirkungen haben. Im Durchschnitt der drei Fächer wird geschätzt, dass ein um eine Standardabweichung höherer Wert auf einer Skala zur Verwendung engagierender Lehrmethoden die Testergebnisse um 4,6 Prozent einer Standardabweichung erhöht. In fächerspezifischen Analysen zeigt sich, dass der Effekt beim Lesen am größten ist.

Mit Blick auf mögliche Handlungsempfehlungen an die Politik, ist darauf hinzuweisen, dass die positiven Effekte bei Schülern mit niedrigem sozio-ökonomischen Status nicht signifikant zu Lasten anderer Schüler gehen. Daher wird der Schluss gezogen, dass eine stärkere Nutzung engagierender Unterrichtsmethoden in der Grundschule – ähnlich wie ein längerer Kindergartenbesuch – ein wirksames Instrument zur Bekämpfung von Bildungsungleichheiten in Deutschland sein kann.

Kapitel 4 befasst sich ebenfalls mit der Grundschulbildung und widmet sich der Schätzung von Klassengrößeneffekten. Die Klassengröße ist einer der wichtigsten Hebel in der Bildungspolitik, da Lehrergehälter in vielen Ländern einen Großteil der Bildungsausgaben ausmachen. Bislang gibt es in der Forschung allerdings keinen Konsens darüber, wie sich Klassengrößenreduzierungen oder -erhöhungen hinsichtlich des Lernerfolgs der Kinder auswirken. Insbesondere lassen sich in quasi-experimentellen Studien häufig nicht vergleichbar große Effekte nachweisen wie in experimentellen Studien. Kapitel 4 ist ein Versuch, diese Forschungsergebnisse in Einklang miteinander zu bringen.

Die Studie zeigt, wie Klassenwiederholungen leistungsschwacher Schüler zu verzerrten Schätzungen von Klassengrößeneffekten führen, denen als Variation Schwankungen in der Jahrgangsgröße innerhalb von Schulen zugrunde liegen. Die Verzerrung wird durch einen negativen Zusammenhang zwischen der ursprünglichen Jahrgangsgröße und dem Anteil sitzengebliebener Schüler in demselben Jahrgang in höheren Klassenstufen hervorgerufen. Sie hat zur Folge, dass die typischerweise negativen Klassengrößeneffekte kleiner geschätzt werden als sie tatsächlich sind. Die Analyse belegt die Existenz eines derartigen Kompositionseffekts anhand von Verwaltungsdaten zu Einschulungszahlen aller Grundschulen im Saarland sowie in Sachsen. Ebenfalls lässt sich zeigen, dass die Verzerrung leicht korrigiert werden kann, indem Kontrollvariablen zu früheren Klassenwiederholungen in das Modell aufgenommen werden. Unter Anwendung dieser Korrektur werden in Kapitel 4 Klassengrößeneffekte auf Grundlage eines Datensatzes geschätzt, der Informationen zu vier kompletten Jahrgängen saarländischer Grundschüler beinhaltet, die am Ende der 3. Klasse an landesweiten Vergleichsarbeiten in Deutsch und Mathematik teilgenommen haben. Die Ergebnisse zeigen, dass sich die Testergebnisse für jeden Schüler weniger in der Klasse um 1,9 beziehungsweise 1,4 Prozent einer Standardabweichung verbessern. Allerdings sind die geschätzten Effekte nicht linear. Während sich die Deutsch- und Mathe-Testergebnisse in größeren Klassen von mehr als 20,5 Schülern für jeden Schüler weniger um 4,8 beziehungsweise 3,8 Prozent einer Standardabweichung verbessern, wird in kleineren Klassen kein Effekt festgestellt. Erhebliche Heterogeneitäten zeigen sich auch in Bezug auf den persönlichen Hintergrund der Schüler. Wie in den anderen Kapiteln auch, profitieren benachteiligte Schüler (solche mit unzureichenden Deutschkenntnissen oder einer Lernbehinderung) überproportional. Schließlich gibt es auch Hinweise auf eine in kleinen Klassen verminderte Wahrscheinlichkeit von Klassenwiederholungen.

Die Ergebnisse von Kapitel 4 liefern erstmals kausale Belege für signifikante Klassengrößeneffekte auf Schülerleistungen in Deutschland. Sie implizieren, dass Reduzierungen der Klassengröße, die darauf abzielen, den Lernfortschritt der Schüler zu erhöhen, auf größere Klassen ausgerichtet sein sollten. Umgekehrt legen sie aber auch Nahe, dass sich die Klassengröße in kleinen Klassen bis zu einem Schwellenwert von etwa 20,5 Schülern je Klasse ohne nachteilige Auswirkungen auf deren Leistungen erhöhen lässt.

Die drei Forschungskapitel dieser Dissertation werden eingerahmt von den Kapiteln 1 und 5. **Kapitel 1** führt in das Thema der kognitiven Fähigkeitsbildung ein und hebt die wichtigsten Beiträge dieser Dissertation hervor. **Kapitel 5** schließt mit einer kritischen Diskussion zentraler Ergebnisse und sowie den daraus folgenden Politikimplikationen.

CHAPTER 1

INTRODUCTION

1.1 Motivation

What is it that makes some people more well-off economically than others? This question has been central to economic debates for as long as economic thought has existed. In fact, it is at the heart of the book "An Inquiry into the nature and causes of the wealth of nations" by Adam Smith, the man who is widely credited to be the founder of modern economics. The same Adam Smith was well ahead of his time by postulating that an individual's abilities and talents constitute a form of capital, much like a machine, that can be used for economic production (Smith, 1776).

The notion of human capital inherent in this thought was only to gain prominence almost two centuries later when the predominant view that capital chiefly consists of physical goods such as machinery, buildings, vehicles, and the like began to be questioned. This questioning was the result of studies on income growth in the US, which found that the growth rate of physical capital possessed by people alone could not account for the observed increase in incomes. Rather, the capital seemed to have been more efficiently deployed (see e.g. Fabricant, 1954; Solow, 1957). As a result, the focus of researchers gradually shifted towards resources such as the knowledge possessed by individuals, their intelligence and the state of health. Much of the popularity that the concept of human capital has since achieved can be attributed to Jacob Mincer (1958), Theodore Schultz (1961), and Gary Becker (1962). Becker (1962) was the first to formulate a unified theory of investment in human capital. He defined human capital as the sum of all physical and mental abilities of people that are relevant for their real income prospects. These abilities are less tangible than traditional physical capital and, since they cannot be traded from one person to another, are inherently hard to measure.

Early empirical work on the importance of human capital for labor market outcomes typically focused on a person's level of education and work experience. These factors were used as predictors of wages in linear regressions; an approach that was invented and popularized Jacob Mincer (1958; 1974), hence the name "Mincerian equations." Mincer showed that in the US every additional year of schooling was associated with a wage increase of more than 10 percent in the late 1950s. While Mincer-type regressions remain popular in labor economics they have been criticized for a number of reasons, most of which revolve around the measurement of human capital. In classical Mincerian equations a person's years of schooling are used as a proxy for that person's human capital. However, in human capital theory as set up by Becker (1962) the process of human capital formation is not limited to the school system. Rather, any activity that embeds resources that are relevant for labor market success in people or improves their physical and mental abilities has a role to play in the formation of human capital. Thus, influences outside the school system such as parents, peers, friends, diet, and hobbies are neglected in simple Mincerian models.[1]

Even if one were to acknowledge that education is not the same as human capital and that estimated returns to the years of schooling and similar metrics such as educational attainment or educational enrollment solely provide information on returns to education instead of human capital, there are a number of drawbacks in connection with measuring education in years. First and foremost, by assigning the same value to every year of education independent of states, countries, institutions and schools, one effectively imposes the assumption on no quality differences in education (Mulligan and Sala-i Martin, 2000; Wößmann, 2003a). This assumptions seems especially problematic in cross-country studies. Most people would probably agree that the added knowledge from one year of education is comparatively larger in rich countries than in most developing countries. One reason for this is that richer countries are able to channel greater financial resources into educational

[1] In fact, the notion of education merely reflecting human capital is closer in spirit to signaling theory as set up by Spence (1973) than human capital theory. Spence (1973) postulates that individuals select an appropriate level of education for themselves depending on their underlying ability to signal this ability to potential employers.

inputs such as teacher training and teaching materials than poorer countries. Second and related to first, the monotonic increase in the education indicator as a result of more time spent in school ignores the fact that faster-learning students often spend less time in the educational system than others (Schneider, 2010). This is easily illustrated in school systems with grade retention. While it is true that retained students are – all else equal – subject to more instruction hours (and years) than non-retained students in their educational careers, they do not receive more instruction content but most likely have to follow the same content twice (or more often depending on the number of repeated grades). Under the plausible assumption that it is possible to master the respective content in the first try (as reflected by the typically large share of non-repeaters), the average quality of instruction for repeaters is lower than for non-retainers due to the higher share of unnecessary repetition. Third, by using a cardinal measure for education, a linearity assumption is imposed that every year in schooling has the same effect on the considered outcome (Wößmann, 2003a). However, if for the completion of a number of tasks on the labor market only a certain threshold of abilities has to be surpassed, the years of schooling that are required to obtain these abilities would have higher labor market returns than subsequent years.

Some of the above-mentioned issue can be remedied. For instance, in order to circumvent the grade retention problem one can use the hypothetical years of schooling that are on average needed to obtain a certain degree. Linearity issues can to some extent be dealt with by decomposing the indicator variable for education into several binary dummies or incorporating higher order polynomials into the regression models. However, the key problem of unobserved quality differences between schools cannot easily be solved. For the sake of cross-country comparisons, Wößmann (2003a) has proposed to weigh years of education by the quality of educational systems. Still, this approach creates a complex set of additional questions that pertain to the measurement of quality differences in education between countries – and possibly between single schools within countries. A first necessary step towards answering these questions is to agree on a set of outcome measures that reflect the goals that different school systems should strive to achieve. Though framed slightly differently, this is very similar to asking how human capital should be measured if not simply by educational attainment or years of education.

Decomposing human capital

A lot of research on specific dimensions of human capital deals with the concept of skills possessed by individuals (Goldin, 2016). To recycle the term used by Becker (1962), skills can be understood as the sum of all mental abilities of a person. Researchers often differentiate between cognitive (intelligence-related) and non-cognitive (personality-related) skills (see e.g. Heckman and Rubinstein, 2001; Heckman et al., 2006; Lindqvist and Vestman, 2011). A distinct advantage of these measures of human capital over conventional schooling indicators is that they can accommodate abilities obtained outside of schools, i.e. in families, from peers and so forth (see e.g. Hanushek and Wößmann, 2008). In the following, the two concepts are briefly reviewed.

First in line are cognitive skills, to which this dissertation is mainly dedicated. While there is no "sharp" definition of cognitive skills – or cognitive ability – in the economic literature, there is agreement that it is linked to the concept of general intelligence. General intelligence, in turn, has been defined by an official taskforce of the American Psychological Association as the "ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" (Neisser et al., 1996, p. 77). In the psychological literature, general intelligence is known as the "*g* factor" and was first discovered by Charles Spearman in 1904 (Spearman, 1904). This factor is derived from the correlations in the share of correct answers of any individual on different cognitive tests. It is based on the observation that a person who does well on one test has a high probability of doing similarly well on any other cognitive test (see e.g. Grabner and Stern, 2011). However, it is widely accepted that cognitive ability is more than "*g*" (see e.g. Carroll, 1993). Frequently, contemporary theoretical models make a distinction between fluid and crystallized intelligence. While the former is closely related to "*g*" and describes the ability to solve complex novel problems, the latter is related to general knowledge on a variety of issues such as vocabulary or general information (Grabner and Stern, 2011).

In empirical research, cognitive ability is usually measured by information on the performance of individuals on general intelligence or knowledge-based tests in fields such as mathematics, science, and reading. Such test data have become widely available over the last couple of decades. This development started in the United States where cognitive skill levels were initially often derived from results on the Armed Forces Qualifications Test (AFQT). The AFQT is a general aptitude test and the primary criterion for suitability for service in the US armed forces (Heckman et al., 2006). Nowadays, large-scale international student assessment studies such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS) enjoy a lot of popularity. Furthermore, household surveys like the German Socio-Economic Panel Study (SOEP) often contain cognitive skill measures such as short IQ tests nowadays (Wagner et al., 2007). Sometimes, researchers also resort to school grades as measures of cognitive skills (see e.g. Müller et al., 2013).

While cognitive skills have been used to approximate human capital for several decades, a more recent development has been the focus on personality traits as predictors of economic outcomes. Starting with work by Heckman and Rubinstein (2001) economists have come to refer to such traits as non-cognitive skills – and, more recently, also as socio-emotional skills (see e.g. Berger et al., 2010; Burgess, 2016). Since human personality is a very complex construct and since, as of yet, no general or even dominant factor for an individual's personality has been discovered in the field of psychometrics that would be equivalent to "*g*" in cognition research, it is unsurprising that the indicators and variables used to capture non-cognitive skills are a lot more diverse than those reflecting cognitive skills (see e.g. Borghans et al., 2008; Thiel and Thomsen, 2013). However, it has repeatedly been shown that different non-cognitive skills are important predictors of a variety of educational, economic, and social outcomes (see e.g. Baron and Cobb-Clark, 2010; Cobb-Clark and Tan, 2011; Goldsmith et al., 1997; Heckman et al., 2006; Heineck and Anger, 2010; Judge et al., 1999; Lindqvist and Vestman, 2011; Lundberg, 2013; Mueller and Plug, 2006; Osborne Groves, 2005). Nevertheless, since non-cognitive skills are not investigated in this dissertation, I will not delve deeper into their measurement and applications in economics here.

Returns to cognitive skills

Any attempt at estimating the returns to cognitive skills has to start by delineating them from other concepts such as non-cognitive skills. Despite the seemingly clear distinction of intelligence-related factors on the one hand and personality traits on the other hand, this task is to some extent hampered by a number of personal characteristics that fall somewhere in the middle. Among others, these include creativity, emotional intelligence, intellectual engagement, and practical intelligence. These factors both provide information on certain aspects of an individual's personality and may influence the results of cognitive tests (Borghans et al., 2008).

Bearing this qualification in mind, a number of studies have been dedicated to estimating returns to cognitive skills. There is ample evidence that cognitive skills are beneficial for a variety of life outcomes such as schooling attainment and wages and even social behaviors such as delinquency and giving out-of-wedlock birth (see e.g. Bowles et al., 2001; Bronars and Oettinger, 2006; Cawley et al., 2001; Green and Riddell, 2003; Hernstein and Murray, 1994). In Germany, Heineck and Anger (2010) provide an analysis based on SOEP data and find positive effects of fluid intelligence on males' wages only, while Dohmen and van Landeghem (2019) find that higher cognitive ability in the form of numeracy skills significantly reduces the duration of unemployment spells. Often, including both direct measures of cognitive skills and years of schooling as predictors in the same model significantly reduces the coefficient of the latter since the two indicators are usually correlated (see e.g. Bronars and Oettinger, 2006; Cawley et al., 2001; Green and Riddell, 2003).[2]

It is important to note that to the extent that cognitive skills are correlated with non-cognitive skills the studies cited above measure a combined effect of the two human capital dimensions. Other studies provide joint evidence by including measures of both dimensions. Heckman et al. (2006) show that both cognitive and non-cognitive skills play important roles in explaining a diverse array of life outcomes. Furthermore, the effects seem to be of similar magnitude. Among the outcomes studied are classical economic variables such as wages but also different (risky) social behaviors such as substance abuse, crime, and teen-

[2] In a cross-country study of economic growth in developing countries, Hanushek and Wößmann (2008) even find that the coefficient of years of schooling becomes zero when a cognitive skill indicator is added to the model.

age pregnancies. Lindqvist and Vestman (2011) reach similar conclusions in a study from Sweden and additionally find that cognitive skills matter more for wages of skilled workers who typically have higher earnings.

Skill formation and main focus of this dissertation

Having underscored the importance of cognitive skills for success in various dimensions of life, it is the aim of this dissertation is to shed light on the *formation* of cognitive skills in children. Knowledge about this process is very valuable to policy-makers as it gives them a lever to raise the level of human capital in a society and as a result set the course for a functioning economy, higher tax revenues, and ultimately higher individual well-being. A necessary precondition for this is that cognitive skills *are* in fact formed during an individual's life and not simply genetically determined. This condition relates to the debate on whether cognitive skills (in particular their general intelligence component) should be regarded as reflecting a person's *innate potential* or rather that person's *realized capacity*. The predominant view on this issue among economists tends to favor the latter, which means that the dichotomy of *nature* versus *nurture* is obsolete in reality. In that sense, abilities are always produced by an interplay of genetic conditions and environmental factors (see e.g. Cunha and Heckman, 2007). These environmental factors may comprise an individual's parental background, educational institutions, his or her professional career as well as other influences such as via peers or relatives (see e.g. Carlsson et al., 2015; Carneiro and Heckman, 2003; Cunha et al., 2006). As argued above, the ability to accommodate all these diverse influences is one of the distinct advantages of direct skill measures over mere schooling indicators.

Next to the existence, amount, and quality of different environmental influences, the *timing* of the exposure to these influences is crucial to understanding cognitive skill outcomes. It has been found that different skills are most easily acquired at different stages of childhood (see e.g. Knudsen et al., 2006). For cognitive skills the so-called critical period seems to be before the age of 10. After this age, at least IQ measures remain fairly stable in individuals (Hopkins and Bracht, 1975). This contrasts with non-cognitive skills, which can also be affected by influences and interventions in adolescence (Cunha et al., 2006). However, on average early influences generate a much higher return than later ones (Cunha and Heckman, 2007). This finding is the result of multiplier effects that are produced by *self-productivity* of skills and *dynamic complementarities*. Self-productivity of skills refers to the fact that skill gaps opened up at one point in childhood persist into higher ages and may even widen because skills are self-reinforcing. Crucially, the concept is not limited to one and same skill. A higher level of non-cognitive skills can also increase cognitive skills and vice versa. For instance, emotional stability (a non-cognitive skill) may strengthen curiosity and explorative drive thereby contributing to the development of cognitive skills (Cunha and Heckman, 2007). Dynamic complementarities are similar to self-productivity and mean that skills acquired earlier in life tend to increase the productivity of investments in skills later on.

The essence of the argument made by Cunha and Heckman (2007) is that interventions that aim at fostering skills in children are best undertaken early. Given the stability of IQ after the age of 10, this should be especially true for cognitive skills. For that reason, this dissertation focuses exclusively on policy interventions that affect children under 10. Importantly, it is dedicated to the production of cognitive skills that takes place in the context of the general school system as well as in the system of universal center-based child care.[3] It is recognized that a multitude of additional factors such as families and peers, extracurricular activities, targeted (preschool) interventions as well as other non-school influences also have a role to play in the formation of cognitive skills. However, an analysis of all these factors would be beyond the scope of this dissertation. This being said, the process of skill formation that takes place in the context of the universal education and child care systems is of particular interest due to the centrality that the objective of skill formation takes up among the goals of these institutions. Furthermore, the education system is probably the most important policy lever for investments in human capital (Burgess, 2016). Against this background, it is the main aim of this dissertation to shed light on the effects of different quantitative and qualitative educational policy levers and come closer to a common understanding what the most important differentiators in school systems with respect to the acquisition of cognitive skills are.

[3] I use the term child care for all forms of center-based day care before the start of primary schooling. Frequently used synonyms in the literature include early childhood education and care (ECEC) as well as preschool.

1.2 Overview and Summary

In total, this dissertation comprises three research articles, each of which is dedicated to one educational policy lever and each of which provides an independent contribution to the literature on cognitive skill formation. Figure 1.1 depicts the educational inputs and cognitive skill outcomes that are analyzed in the three research articles. In addition, it illustrates when in the lifecycle of a child the interventions take place and when outcomes are measured. Table 1.1 provides a detailed overview of all chapters including their research question(s), the main findings, the data and methodology that is used as well as collaborations with co-authors. In the following, the three chapters are briefly summarized.

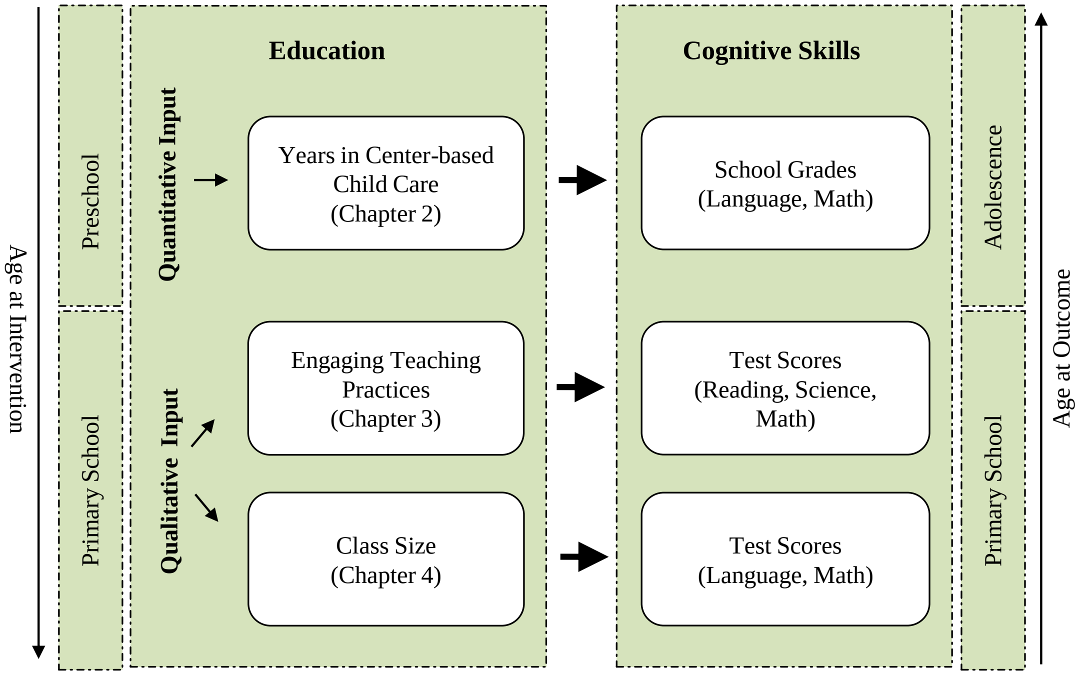
Chapter 2 focuses on medium- and long-run effects of universal center-based child care attendance. It exploits an amendment to the “Child and Youth Welfare Act” (*Kinder- und Jugendhilfegesetz*) in 1992 that established the legal right to a heavily subsidized half-day place in child care for all children from the age of 3 until the beginning of primary school by 1996. The reform took place against the background of severe demand rationing in the Western part of the country that could not quickly be alleviated despite a rapid expansion in child care supply. Hardest hit by the undersupply were 3-year-olds as places were often assigned by age.

Exploiting the fact that the expansion in child care supply was staggered across counties for arguably exogenous reasons, this chapter analyzes the causal effects of an additional year of center-based child care attendance on a variety of education-related outcomes in adolescence in an instrumental variables framework. Practically, administrative information on the county-level slot-child-ratio among 3- to 6.5-year-olds is used as the excluded instrument that reflects the level of regional child care supply. The cognitive skill indicator employed in this study is school grades in German language and mathematics. Next to cognitive skill measures the chapter includes several other outcomes measures in order to get a comprehensive picture of the effects of longer child care attendance. These measures include secondary school track choice, grade retentions, and future educational aspirations. All outcomes are measured when children are 16 or 17 years old and first enter the SOEP, a large representative household survey on which the analysis is based. In addition to complete annual data on children's child care careers, the SOEP provides rich background information on children and their families.

The results indicate that German language grades are positively influenced by longer child care attendance. The effect is especially pronounced among weaker students as reflected by a sizably reduced likelihood of obtaining one of the three worst grades. There is also evidence for increased educational aspirations, again most notably at the lower margin where students either aspire to a vocational degree or no further degree after leaving the general school system. No effects are found for secondary track choice and the probability of ever having repeated a grade. Taken together, the results provide additional evidence for the well-established finding that child care attendance is most beneficial for disadvantaged children. The study complements the literature on medium- and long-run effects of center-based child care by focusing on a wide array of outcomes measures, most of which have not been causally analyzed before in the German context. With regard to policy implications, the case for public funding of child care centers as a means to tackle inequality in child opportunities is strengthened. This effect comes on top of the well-known positive effects of center-based child care on maternal labor supply (for maternal employment effects of the same reform that is investigated in Chapter 2, see Bauernschuster and Schlotter, 2015).

Chapter 3 takes the analysis a step further in the lifecycle to the period of primary schooling. It deals with the question what classroom actions by teachers are effective in conferring cognitive skills upon students. This question is of central importance to policymakers, school principals, and teachers who look for ways to maximize educational output. Furthermore, analyzing what teachers do in classrooms and how they interact with their students is of particular relevance, since socio-economic teacher characteristics such as gender, experience, and education cannot account for the huge achievement differences attributable to different instructors (see e.g. Lavy, 2015). Specifically, the chapter assesses the effectiveness of employing teaching practices that potentially increase the students' engagement with the course content. These practices include summarizing key messages, relating lessons to students' daily lives, use questioning, encouraging students, giving praise, and bringing interesting materials to class. The study thereby goes beyond the traditional dichotomy of "modern" versus "traditional" teaching that has dominated the economic literature on teaching methods.

Figure 1.1: Structure of dissertation



Age at Intervention
Preschool
Primary School
Education
Quantitative Input
Qualitative Input
Years in Center-based Child Care (Chapter 2)
Engaging Teaching Practices (Chapter 3)
Class Size (Chapter 4)
Cognitive Skills
School Grades (Language, Math)
Test Scores (Reading, Science, Math)
Test Scores (Language, Math)
Adolescence
Primary School
Age at Outcome

Source: Own illustration.

The analysis is based on a unique dataset that combines information from the 2011 waves of the TIMSS and PIRLS studies for a representative sample of German fourth-graders. TIMSS is an achievement study that deals with mathematics and science, while PIRLS is dedicated to reading. Therefore, test scores are observed in three different subjects for each student. This allows the use of within-student estimation for identification. The results indicate that engaging teaching practices yield non-negligible achievement gains among students from low socio-economic backgrounds. It is estimated that a one-standard-deviation-increase on a composite scale measuring the use of potentially engaging teaching practices raises test scores in math, science and reading by 4.6 percent of a standard deviation of the test score distribution. Subject-specific analyses suggest that the association is strongest in reading. These benefits do not come at the cost of significantly lower achievement among students from high socio-economic backgrounds. Similar to the findings of Chapter 2, the results of this chapter open up a potential avenue for tackling inequality in Germany.

Chapter 4 is also concerned with primary schooling and dedicated to one of the central questions in the economics of education, namely whether smaller classes lead to higher student achievement. The fact that teachers' salaries account for the bulk of educational spending in most countries alone makes class size one of the key policy levers in school systems. This notwithstanding, there is no consensus among researchers about the effects that class size reductions or increases generally have. Chapter 4 is an attempt at reconciling some of the mixed evidence between experimental and quasi-experimental studies. Typically, class size effects are much smaller in the latter kind of analyses.

In the study we theoretically illustrate that grade retentions of poorly performing students give rise to an upward bias in class size estimates based on within-school variation in cohort size over time. This bias is the result of a mechanical relationship between initial cohort size and the share of previously retained students in the same cohort in higher grades. The existence of such a compositional effect finds empirical support in administrative data on school enrollment for all primary schools in the German state of Saarland. The resulting bias can be easily corrected by controlling for whether or not a student has previously been held back a grade. We perform this correction and estimate class size effects utilizing data that covers four entire cohorts of students in Saarland that participated in state-wide centralized exams in German language and mathematics at the end of grade 3. Analogous to Chapter 3, this chapter therefore employs test scores in different subjects in primary school as cognitive skill indicators. Instrumenting class size in grade 3 by predicted class size based on imputed cohort size, we find that test scores are increased by around 1.9 and 1.4 percent of a standard deviation for each one-student decrease in class size. In line with the theoretical predictions, this effect is considerably larger than without controlling for previous retentions. What is more, we find evidence for a decreased likelihood of grade repetitions in smaller classes. However, class size effects seem to be highly non-linear. Whereas language and math test scores are increased by 4.8 and 3.8 percent of a standard deviation in larger

Table 1.1: Overview of chapters

	Chapter 2	Chapter 3	Chapter 4
Title	The Effects of Universal Child Care Provision in Adolescence	Engaging Teaching Practices and Achievement – A Within-Student Approach in Three Subjects	Birth Cohort Variation and the Estimation of Class Size Effects
Research question(s)	What effects does an additional year in center-based child care at the age of 3 have on a variety of schooling-related outcomes in adolescence?	What effects does the use of potentially engaging teaching practices in primary school have on test scores in the short-run?	What effects does class size in primary school have on test scores in the short-run? How does initial cohort size affect class size estimates based on cohort size variation?
Main finding	Longer center-based child care attendance improves school grades particularly among weaker students and increases future aspirations towards obtaining a vocational degree; No effects are found for secondary school track choice and the likelihood of grade retentions	The use of potentially engaging teaching practices improves achievement only among students from low socio-economic backgrounds	Smaller classes improve achievement in language and math and reduce grade repetitions; Failure to control for previous grade retentions results in an upward bias in class size estimates based on cohort size variation
Data	Survey data (SOEP) + administrative data (Federal Statistical Office)	Survey data (TIMSS & PIRLS)	Administrative data (SOE, Statistical Offices of Saarland and Saxony) + survey data (NEPS)
Identification method	Instrumental Variables Estimation	Student Fixed Effects Estimation, Correlated Random Effects Estimation	Instrumental Variables Estimation
Co-author	-	-	Maximilian Bach

Source: Own illustration.

classes of more than 20.5 students for each student less, we fail to find an effect in smaller classes. Significant heterogeneities also appear with regard to student background. Again, disadvantaged students (in this case students with a migration background, insufficient German proficiency, a learning disability, or who have been retained in the past) benefit disproportionately strongly.

Chapter 4 offers first causal evidence of significant class size effects on test scores in one federal state of Germany, a country where educational researchers have been particularly vocal in disputing the merits of class size reductions in the past. Policy implications can easily be drawn. First of all, class size reductions to increase student achievement should be targeted at larger classes. Conversely, class size in small classes may even be increased up to a certain threshold without negative consequences for student achievement.

Finally, Chapter 5 discusses the general findings of this dissertation and hints at possible avenues for future research.

1.3 Common Contributions

While the three main chapters of this dissertation make independent contributions to the economics of education literature, there are a number of recurrent themes that link the articles and can therefore be considered as common contributions.

The *first and central* common contribution of the three chapters is their focus on the *formation of cognitive skills*. More precisely, all chapters deal with *cognitive skill returns to interventions in education systems*. Cognitive skill acquisition can be classified as an intangible effect of education in the sense that such skills neither entail any direct monetary rewards nor make a direct statement about success on the labor market (see e.g. Dahmann, 2016). Since economists have traditionally been interested in tangible, monetary outcomes of education, the literature on cognitive skill returns is still quite patchy and leaves a number of research questions unanswered. This dissertation is dedicated to three of them, namely the long-run effects of center-based child care attendance, the short-term effects of applying potentially engaging teaching practices in primary school, and the short-term effects of class size in primary school. By focusing on such a diverse array of interventions that include both quantitative (longer child care attendance) as well as qualitative inputs in education (teaching practices and class size), this dissertation acknowledges the complexity of educational policy.

A *second*, related contribution is the *focus on the first half of childhood*, i.e. the first ten years of a child's life. This focus is necessary, as general cognition has been found to be particularly malleable at this age (Cunha and Heckman, 2007). While crystallized intelligence may be acquired later on, fluid intelligence is rather stable after the age of ten (Hopkins and Bracht, 1975). This means that any intervention aimed at increasing fluid intelligence is not only less efficient later on in the life cycle (as is the case with non-cognitive skills), but often outright ineffective.

The *third* common contribution is the fact that all chapters use data from Germany. This in itself is an advantage, as the effects of different interventions can be compared to each other and therefore provide decision-makers with a more comprehensive picture of their policy options. Furthermore, the general education system as well as the system of universal child care boast a number of specificities that complement the many studies in the field that originate from either the USA, the United Kingdom, or Scandinavia. Most prominent among these specificities is the fact that child care is heavily subsidized and therefore inexpensive as well as the fact that the school system is heavily tracked starting in secondary school. The latter feature is especially relevant for long-run studies such as conducted in Chapter 2.

A *fourth* common contribution is of methodological nature and pertains to the attempt at identifying *causal effects*. Being able to establish causality between a reform (a treatment) and an outcome is crucial for policy purposes, as it gives decision-makers the maximum amount of information on what to expect from their actions. The "gold standard" towards reaching this goal is to conduct carefully planned experiments (randomized controlled trials). By randomly assigning the treatment to a subgroup of individuals out of all participants in the experiment, it is possible to compare outcomes between the two groups that should not be different in any characteristic except their treatment status. This way, the problem of the missing counterfactual, i.e. that one and the same person cannot be observed both as treated and as untreated, is solved. However, in reality it is often not possible to conduct such experiments. This may have to do with practical reasons (e.g. lack of funding) or ethical reasons (see e.g. Athey and Imbens, 2017). Therefore, researchers routinely resort to natural (or quasi-) experiments (for an overview of often-employed quasi-experimental strategies, see Angrist and Pischke, 2009). Such quasi-experiments are characterized by the fact that, while not intended as experiments, treatment is still randomly assigned due to some specific feature of the reform or the setting in which the analysis takes place. In Chapter 2, this feature is the place of residence of the child's family. Since child care supply was higher in some places as opposed to others for arguably exogenous reasons, children living in areas with higher child care supply had a higher likelihood of entering child care early than others. This mechanism is exploited for identification in an *instrumental variables framework (IV)*. In Chapter 3, I use *fixed effects* and *correlated random effects estimation* in a within-student between-subjects framework for identification. Instead of comparing treated and untreated individuals, I am here comparing the same student's performance in different subjects and relate it to teachers' instructional practices. Based on some assumptions, this approach tries to mimic the (unattainable) ideal of observing the same individual in different treatment statuses at the same time. Finally, in Chapter 4 another *IV approach* is employed. This time, the size of the cohort a child was born into serves as the exogenous feature that influences the likelihood of ending up in a larger or in a smaller class in primary school.

The *fifth* common contribution is the combination of *different data sources* that complement each other in the same study. As a rule, researchers try to exploit the kind of data that are most suited to answering the research question at hand. However, oftentimes no ideal dataset is available that caters to all empirical needs. For example, while administrative data often contain huge numbers of observations, they frequently suffer from limited background information on each individual. Furthermore, they usually do not contain subjective information on individuals, for instance on their future aspirations. On the contrary, survey data often provide rich sets of control variables but have the drawback of limited sample sizes. By merging different data sources or performing analyses on different datasets it is often possible to get the best out of different worlds. In Chapter 2, I merge administrative data on child care supply at the county level to survey data from the SOEP. In Chapter 3, data from two different surveys are merged, namely the 2011 waves of the TIMSS and PIRLS studies. In Chapter 4, administrative data on school enrollment is merged to an extraordinarily rich dataset of test scores for the full population of third-graders in the German state of Saarland. What is more, data from the German National Educational Panel Study (NEPS) as well as administrative data on enrollment and grade retentions in the state of

Saxony are used to verify some of the predictions of the theoretical model that would not have been possible with the test score dataset from Saarland.

Finally, a *sixth* common contribution of all studies is their strong emphasis on *effect heterogeneities*. These heterogeneities can pertain to differential effects (or effect sizes) on different subgroups of the population or non-linear effects along the distribution of the main explanatory variable. As for population subgroups, all chapters separately estimate effects on boys and girls as well as students from different socio-economic backgrounds. Since there is no universally agreed indicator for socio-economic background, different measures that are frequently encountered in the literature are employed in different chapters. In Chapters 2 and 3, socio-economic background is determined by the level of education of the mother and both parents, respectively, while in Chapter 4 the number of books at home is used. Similarly, there are different ways of uncovering effect non-linearities. In Chapter 2, treatment dummies that split the linear treatment indicator on school grades into different segments are considered. In Chapter 3, a squared term of the treatment variable is added to the models, whereas in Chapter 4 spline regressions are carried out.

CHAPTER 2

THE EFFECTS OF UNIVERSAL CHILD CARE PROVISION IN ADOLESCENCE

2.1 Introduction

In recent years, several industrialized countries have significantly expanded their supply of publicly funded child care. One objective is to improve the possibilities for young parents to reconcile work and family life. Another objective is to positively influence the children's own life prospects.[4] Against this background, it is highly relevant to investigate the effects of universal child care provision on skill-related outcomes that are relevant for a child's success in life.[5] From an economic perspective, the effects of child care attendance should be evaluated over the medium- and long-run.[6] The reason for this is that outcomes in adolescence and young adulthood are more directly relevant for professional success than short-run outcomes since they are measured at a time when individuals start thinking about their working lives. In fact, many factors that determine lifecycle incomes are already in place by this time (see e.g. Cunha et al., 2006).

[4] For discussions on the aims of German child care policies, consider Koebe and Spieß (2019) and Spieß (2011).

[5] I use the term "universal child care" for all forms of center-based child care that are generally open to all children. The opposite of "universal child care" would be "targeted child care" that aims at certain − often disadvantaged − groups of children.

[6] Roughly speaking, short-run outcomes can be observed until the end of primary school, medium-run outcomes revolve around the secondary school choice as well as attainment in lower secondary school, and long-run effects essentially comprise all outcomes from the end of compulsory schooling at around age 16 onwards. The medium-run effects and the long-run effects are grouped since there is no agreement on the exact age threshold separating one from the other.

The focus of this study is on medium- to long-run effects of center-based child care attendance on educational and skill outcomes as well as aspirations towards higher secondary or post-secondary education. There are a number of theoretical reasons why one would expect center-based child care to affect child development. First, as public care is substituted for home care the quality of caregiving may increase, decrease, or remain stable (see e.g. Spieß, 2017). Second and related to first, likely increased interactions with other children in child care centers may play a role in skill and personality formation. Third, increased exposure to away-from-home care may feed back into the home environment and thereby affect the quality of home care (see e.g. Kuger et al., 2019). Fourth, the replacement of home care frees up time for other activities by the parents such as market work that increases family income, which can be invested to support child development (see e.g. Spieß, 2017). It is a priori unclear if the net effect of center-based child care is positive, zero or even negative, since it directly depends on the relative quality of care at home and in publicly funded child care centers. Since the quality of home care differs between different groups of society, it is likely that there are differential effects on different children, with children from low socio-economic backgrounds theoretically standing to gain the most (see e.g. Knudsen et al., 2006). There is agreement among economists that any intervention that aims at conferring skills upon children is best undertaken early, as existing skills facilitate the acquisition of additional skills (see e.g. Cunha and Heckman, 2007; Heckman and Masterov, 2007). This is especially important for children with low skill levels who, in the absence of effective support, would fall ever further behind other children.

To investigate child care effects, I draw on the German Socio-Economic Panel Study (SOEP), an extensive household survey that started in 1984. As of 2018 there are nearly 15,000 participating households (Goebel et al., 2018). The SOEP provides complete annual information on children's child care careers as well as rich background information on children and their families. Outcomes under study include information on educational trajectories (secondary track choice, grade repetitions), cognitive skills (school grades),[7] as well as aspirations towards further education. All outcomes are measures when children are 16 or

[7] I am aware of the fact that school grades usually measure more than "pure" cognitive skill levels. Rather, they are the result of cognitive skills and more non-cognitive skill-related traits such as motivation, discipline, and engagement in class. Bearing this in mind, I still refer to school grades as cognitive skill measures, thereby following other researchers in the field (see e.g. Müller et al., 2013).

17 years old and first enter the SOEP as respondents, although track choices and grade repetitions refer to events that have happened in the past.

In order to estimate the desired effects I adopt an instrumental variables (IV) framework that takes care of endogenous sorting of children into child care along unobserved factors. I use information on the child care slot-child-ratio of 3- to 6.5-year-olds at the county-level as the excluded instrument. This ratio varied drastically in West Germany at the time under study which is the late 1980s and the 1990s. Crucially, the respective ratios did not reflect local market-clearing equilibriums, since demand was severely rationed virtually everywhere as reflected by long waiting lists. Since child care places were often assigned by age, rationing was most severe for 3-year-olds, whose attendance rate in 1995 was only 30 percent as compared to 60 percent among 4-year-olds and 90 percent among 5- and 6-year olds (Bauernschuster and Schlotter, 2015; Cornelissen et al., 2018). Against this background, the federal government amended the "Child and Youth Welfare Act" (*Kinder- und Jugendhilfegesetz* [KJHG]) in 1992, stipulating that all 3-year-olds and over would have a legal right to a heavily subsidized place in child care by 1996. This led to a rapid and staggered expansion of child care facilities that, however, varied across counties due to different administrative and financial constraints (see e.g. Cornelissen et al., 2018; Kreyenfeld et al., 2000). As a result, a lot of variation in regional child care supply emerged for plausibly exogenous reasons. In practice this variation meant that depending on their place of residence some parents found it harder to secure a place in child care for their offspring than others. I support the exogeneity of regional child care supply by conditioning on a wide range of key determinants of child care demand and by showing that supply is unrelated to parental personality measures not included in the main analyses. Further, I demonstrate that supply levels are orthogonal to mobility rates, thereby ruling out endogenous migration into counties with higher supply.

The analysis is facilitated by the fact that in response to municipalities' difficulties in meeting the target of universal child care supply for all children aged 3 or older, the German Parliament passed legislation that allowed municipalities to introduce day-of-birth cut-off rules on who would be eligible for child care and who would not.[8] Usually this cut-off was the start of the school year in either August or September depending on the federal state and year. Consequently, children born after the cut-off often could not enter child care in the year that they turned 3 but had to wait for another year. This renders the annual information on child care attendance provided by the parents into a good proxy for actual child care attendance of one year. In the analyses, those children who entered child care at the beginning of the first school year after their third birthday are considered treated while those who enter child care one year later are the control group. Relating the treatment status to regional child care supply levels gives the local average treatment effect (LATE) of one additional year of out-of-home child care for the group of *compliers*, i.e. those who enter into treatment solely because child care supply is higher in their county than elsewhere. This group of *compliers* comprises children from families with a low resistance to center-based child care who typically hail from more advantaged backgrounds, since child care take-up follows a social gradient in Germany (Bach et al., 2019; Cornelissen et al., 2018; Felfe and Lalive, 2018; Jessen et al., 2019; Kühnle and Oberfichtner, 2017; Schober and Spieß, 2013; Schober and Stahl, 2014; Scholz et al., 2019). It is, therefore, possible that other groups of children, i.e. those with a lower quality care environment at home, may be differentially and perhaps more positively influenced by child care attendance.

The results indicate that German language grades are positively influenced by an additional year in child care. Particularly, I find a reduced likelihood of obtaining bad grades. Longer child care attendance also increases educational aspirations in some specifications. Again, it is particularly at the lower margin, namely the aspirations towards a vocational degree as compared to no degree, where an additional year of child care has beneficial effects. These findings are in line with large parts of the literature that theoretically postulate and empirically prove that child care effects are most marked among disadvantaged children. No effects are found for secondary track choice and grade repetitions as well as aspirations towards tertiary education. When interpreting the coefficients, which are quite sizeable, one should bear in mind that due to the limited sample size precision is compromised in some instances. The results are therefore best understood as providing guidance on

[8] Concretely, the right to use cut-off rules was stipulated in the Second Amendment to Volume 8 of the Social Code [Zweites Gesetz zur Änderung des Achten Buches Sozialgesetzbuch] on 21 December 1995.

where to look for significant effects and what signs to expect rather than pinpointing exact effect sizes. In this light, they prove that longer child care attendance should in no case be detrimental to child development even among early "takers" of publicly funded child care and may have positive medium- and long-run effects on skills as well as future educational aspirations.

This study complements the literature on medium- and long-run effects of center-based child care by investigating its impact on a wide array of outcomes, most of which that have not been causally studied in Germany before. This is particularly true for school grades and educational aspirations. The chapter further contributes to the literature on track choices in a heavily tracked school system such as the German one, employing a novel estimation strategy and exploiting a rich dataset that has previously not been used for this purpose.[9]

The remainder of this chapter is structured as follows: Section 2.2 provides a brief overview over the existing literature on the topic. Section 2.3 introduces the institutional features of the German child care system. Section 2.4 presents the empirical strategy, followed by an outline of the data in section 2.5. Section 2.6 presents the results. Section 2.7 concludes.

2.2 Related Literature

There is a vast literature on the effects of center-based child care attendance. A large part of this literature deals with interventions targeted at particularly vulnerable children from low socio-economic backgrounds.[10] The effects of these interventions are not directly compara-

[9] In their study on children's medium-run cognitive and non-cognitive skill outcomes that include track choice, Kühnle and Oberfichtner (2017) use data from the German National Education Panel Study (NEPS). Estimating fuzzy regression discontinuity models, they, too, fail to find any effects on track choices.

[10] The largest of these programs that are mostly geared towards disadvantaged children is Head Start, a federal program that has served some 31 million children in the United States since its founding in 1965 (Head Start, 2013). The evidence on it is mixed. While significant short-run improvements in literacy, language and maths as well as emotional maturity could be established after only one year of program exposure in randomized trials, these effects were rather small in size and often vanished soon after (Barnett, 2011; Puma et al., 2006; Vogel et al., 2010). However, Garces et al. (2002) as well as Ludwig and Miller (2007) establish tentative evidence for higher high school completion rates while Carneiro and Ginja (2014) provide empirical support for lowered behavioral and health problems. Among the most well-known smaller and more intensive programs are the High Scope Perry Preschool Program, the Carolina Abecedarian Early Intervention Program, the Early Training Project as well as the Milwaukee Project (for an overview of these programs see Currie, 2001). All these programs had a strong and significant positive impact on scholastic success, which, however, in many

ble to those of universally available child care as they often employ a highly intensive care mode and address children with likely worse quality of care at home. Nevertheless, there is a growing evidence base on universally available child care as well. The empirical results these studies have produced are rather mixed, which is unsurprising since the studies differ in terms of (a) at what age children are treated, (b) at what age outcomes are measured, (c) which outcomes are analyzed, (d) which countries and institutional features are studied, and (e) what identification strategy is employed. However, a general conclusion to be drawn is that positive effects of child care attendance at young ages on a variety of cognitive and non-cognitive skill indicators are mostly found for children from low socio-economic backgrounds (overviews of the literature on universal child care can be found in Baker, 2011; Dietrichson et al., 2018; Ruhm and Waldfogel, 2011; Schlotter and Wößmann, 2010).

Most closely related to the present study are papers dealing with medium- and long-run effects of child care attendance, in particular those that focus on child care spells in preschool age, i.e. between the ages of three and six.[11] Among the studies from Germany, Bach et al. (2019), Müller et al., 2013), Kühnle and Oberfichtner (2017), and Schlotter (2011) are most relevant. Bach et al. (2019) study the acquisition of non-cognitive skills as a result of longer child care attendance. For identification, they use a very similar setup including an IV approach exploiting regional variation in child care supply based on the same policy reform as in the present paper. They find significant positive effects of an additional year in child care on the personality trait of extroversion as well as to some extent on openness in adolescence. Müller et al. (2013) study school grades in adolescence and find correlations between longer child care attendance and better grades that, however, vanish in sibling models based on family fixed effects. They further study secondary track choice and find a decreased likelihood of attending the lowest track (*Hauptschule*) when child care is attend-

cases was stronger in the short-term then in the long-term (literature reviews are provided by Chambers et al., 2010; Crane and Barg, 2003; Currie, 2001; Yoshizawa et al., 2013). Furthermore, most authors find that effects on school-related outcomes are larger or last longer for females than for males (see e.g. Anderson, 2008; Barnett et al., 1998; Campbell et al. 2002; Heckman et al., 2013; Sandner, 2013).

[11] There is a rather small literature on the effects of child care attendance at very young ages, i.e. between zero and two. The findings of this strand of the literature on early childhood interventions are mixed. While some studies that are mostly dealing with countries in which quality of center-based child care is rather low find negative effects on child outcomes in the short- and long-run (see e.g. Baker et al., 2008; Herbst, 2013; Fort et al., 2017), others find positive effects (Datta Gupta and Simonsen, 2010; Drange and Havnes, 2014; Noboa Hidalgo and Urzúa, 2012). In Germany, Felfe and Lalive (2013; 2018) find positive effects on general child development and school readiness that disproportionately accrue to boys and children from disadvantaged families in marginal treatment effects analyses.

ed for two years instead of just one. For all longer attendance durations, no effects are found. Kühnle and Oberfichtner (2017) fail to find medium-run effects on cognitive and non-cognitive skills as well as secondary track choice due to an increase in the duration of child care attendance of about four months. They employ a fuzzy regression discontinuity design exploiting the fact that many children born in the last quarter of the calendar year enter child care at the start of the school year in which they turn three, i.e. before they become eligible via their third birthday, while children born at the beginning of the next year often wait until the next summer. They thereby corroborate earlier research results by Schlotter (2011) who fails to find an effect on secondary track choice in sibling models using family fixed effects. As for short-run effects in Germany, Cornelissen et al. (2018) find that children who have attended child care longer score better on primary school entry examinations than others in a marginal treatment effects framework. Gains are largest for those who are least likely to attend due to their worse alternative outcomes, for instance immigrant children.

Studies from other countries come to similarly mixed conclusions on medium- and long-run effects (see e.g. Dietrichson et al., 2018). There are a number of studies that find long-lasting benefits that in some instances reach well into adulthood. Among these, two are particularly relevant for the sake of this paper, as they also exploit regional variation in child care access for identification and consider outcomes in adolescence. These are Datta Gupta and Simonsen (2016) and Dumas and Lefranc (2012). Datta Gupta and Simonsen (2016) find positive effects of center-based child care attendance at age 2 on language grades in a sample of Danish ninth graders. Dumas and Lefranc (2012) estimate that the likelihood of grade repetitions is reduced and the probability of graduating from high school increased due to child care attendance at the ages of 3 and 4 in France. Another study that yields significant positive estimates and is relevant for the sake of this paper is Apps et al. (2013) who look at a multitude of different outcomes in adolescence and young adulthood in England that include intentions towards further education. They find positive effects on these intentions that are especially pronounced among children from low socio-economic backgrounds. However, their results may suffer from endogeneity bias as they only condition on observables in a matching framework.[12] In addition to these studies that find positive mean effects, there are a number of additional studies that find positive effects in adolescence only among certain subgroups, typically children from disadvantaged backgrounds. Examples for these studies are Cascio (2009), Cascio and Schanzenbach (2012) as well as Smith (2015) in the US and Felfe et al. (2015) in Spain. However, there are also a number of studies that fail to find significant effect in adolescence and in some instances estimate rather precise nulls. These include Drange et al. (2016) who study end-of-school exams, high school drop-out and academic track in Norway as well as Blanden et al. (2016) who focus on test scores of 11-year-olds in the context of a private sector expansion of child care in England. Finally, DeCicca and Smith (2013) even find negative effects for starting kindergarten one year earlier in Canada in terms of tenth-grade math and reading scores. Looking at outcomes later on in life, Havnes and Mogstad (2011) provide evidence for positive effects on educational attainment, labor force participation and less welfare dependency of individuals in their 30s in Norway. Children with low-educated mothers and girls benefit most. In a later analysis of the same child care reform that led to a large-scale expansion of subsidized child care, Havnes and Mogstad (2015) establish that effects were only positive in the lower and middle part of the income distribution and even negative in the upper part. Similarly, Herbst (2017) finds positive long-term employment effects that disproportionately accrue to the most economically disadvantaged.[13]

Possibly, some of the seemingly contradictory evidence can be reconciled by differences in the quality of child care provision. For instance, Bauchmüller et al. (2014) demonstrate that long-run effects on cognitive development crucially depend on quality indicators such as the number of staff per child as well as the gender and education background of staff in Denmark. In a recent study from Germany, Camehl (2018) shows that high quality child care has a small positive impact on children's non-cognitive skills by exploiting with-

[12] Further studies using similar research designs are Goodman and Siamesi (2005) as well as Fessler and Schneebaum (2016) who both find positive effects on a variety of different long-term outcomes that include the likelihood of obtaining an educational degree, wages, and working full-time in England and Austria, respectively.

[13] There are a couple of additional studies that provide evidence on child care effects in low-income countries. Bietenbeck et al. (2018) provide positive medium-run evidence on school progression and test scores in Kenya and Tanzania, while Bastos et al. (2017) corroborate the positive findings in terms of school progression in rural communities of Guatemala.

in-center differences in a variety of quality indicators. Ambiguous results may also be the cause of different alternative modes of care, namely parental care or informal non-parental care. The latter of the two is often associated with lower quality which should lead to especially large benefits when it is replaced by formal child care (see e.g. Danzer et al., 2017).

2.3 Institutional Background

Child care in Germany is part of the child and youth welfare system. Due to the federal nature of the German state, different levels of government are involved in it. The federal government has legislative and organizational powers through its authority for public welfare. The states (*Länder*) are responsible for the implementation of legislative acts, while the municipalities (*Gemeinden*) have to ensure the actual provision of services and share the funding with the states. This division of tasks mirrors the principle of subsidiarity, a fundamental paradigm in German policy-making, which stipulates that societal services should be delivered at the lowest possible social unit. Consequently, almost all child care centers are either operated by municipalities or by licensed non-profit providers that are funded by municipalities. Most prominent among non-profit organizations are church-related providers such as *Caritas* or *Diakonie* as well as other large welfare providers such as the *Paritätische Wohlfahrtsverband* and the *Arbeiterwohlfahrt.* Since the system is heavily regulated in terms of quality of staff, child-staff-ratios, and construction norms among others and an official license is required to enter the market, no noticeable private for-profit child care sector has developed (Spieß, 2008). As a result of the subsidies, parents shoulder only about 10 percent of child care costs (Cornelissen et al., 2018).

During the period under study, kindergarten was usually part-time so that the children would be home again for lunch (OECD, 2004). Quality was partly ensured by state-specific standards on issues such as group sizes, opening hours, staff-child ratios, and space (Spieß, 2008). Generally, children in Germany are supposed to be supported in their development through play and informal learning. This contrasts with other countries like France where more emphasis is put on structured and formalized forms of learning already in child care (Chartier and Geneix, 2007). The German understanding of child care is rooted in the social pedagogy tradition (*Sozialpädagogik*), which can be described as a holistic approach to learning, caring and upbringing instead of a narrower focus on some of these areas. The importance of the concept is exemplified by the fact that child care workers in Germany routinely identify themselves as “pedagogues” rather than “teachers” (OECD, 2004).

2.3.1 The Expansion of Child Care Supply

Demand for child care exceeded supply in most West German municipalities in the early 1990s. Consequently, social planners had to find ways of how to allocate slots to applicants. Often, available places were granted by age so that 3-year-olds were more affected by demand rationing than 4-year-olds, while 4-year-olds were more affected than 5-year-olds, and so on. Another relevant factor was the mother’s employment status. Children with working mothers had better chances to secure a slot than others. To further be able to differentiate between children of the same age and with the same maternal employment status, most child care centers operated waiting lists (Cornelissen et al., 2018).

Against this background, a public debate on child care provision quickly evolved in the newly reunified Germany. This debate led to an amendment of the “Child and Youth Welfare Act” (*Kinder- und Jugendhilfegesetz*) in 1992 that laid down a legal claim to a part-time kindergarten slot for all children aged 3 or older until the start of primary school. This claim was to enter into force by 1 January 1996. The passing of the law change led to a drastic expansion of child care supply that, however, varied starkly between different municipalities due to different financial and administrative constraints (see e.g. Cornelissen et al., 2018; Kreyenfeld et al., 2000). It quickly became obvious that many municipalities were over-burdened with the requirement of ensuring child care slots to all children between 3 and school-starting age. Due to the decentralized nature of the system and the staggered expansion of supply, large differences in kindergarten slots existed in the mid-1990s not only between East and West but also within West Germany. This is best exemplified by the level of supply in different counties (*Landkreise und kreisfreie Städte*), which are federal entities that are smaller than states but usually contain several municipalities.[14] For instance, while there was less than one slot for two children between 3 and 6.5 years in the county of

[14] The exemption to this rule are so-called county-level cities (*kreisfreie Städte*), in which case county and municipality are the same. In total, there are slightly more than 400 counties in Germany.

Aurich in Lower Saxony in 1994, other counties such as Cochem-Zell in Rhineland-Palatinate and Baden-Baden in Baden-Wurttemberg provided roughly one slot for every child of the same age group. Similarly, in 1998 the county of Wittmund in Lower Saxony still only provided 49.7 child care slots for every 100 children between 3 and 6.5 while many counties in Baden-Wurttemberg and Rhineland-Palatinate had completed the required expansion (Statistisches Bundesamt, 2013). In response to many municipalities' difficulties in providing sufficient places, they were granted the right to operate cut-off rules until the end of 1998. Specifically, municipalities could use the start of the school year (usually in August or September) as a cut-off point. Children who had turned 3 before the cut-off were given a child care slot while children who would turn 3 after it would have to wait for another year. That way, counties were able to streamline entry into child care with the start of the school year, an option that was heavily used as Bach et al. (2019) and Schlotter (2011) are able to demonstrate empirically.[15]

2.4 Empirical Strategy

Exploiting the under-provision of child care places in the late 1980s and 1990s and the transitory cut-off rules that allowed municipalities to streamline acceptance into child care with the start of the school year, the sample can be divided into children who started child care one year earlier than other children. The first group comprises children who were assigned a child care slot at the start of the first school year after their third birthday while the second group comprises all those who entered one year later. From here on, I will refer to the first group as the treatment group and to the second group as the control group. Relating treatment status to the respective outcome of interest while conditioning on a set of covariates gives the following model:

$$Y_{si} = \alpha_s + \beta_s T_i + \gamma_s X_i + \varepsilon_{si}, \tag{2.1}$$

[15] Bach et al. (2019) use data from the German National Educational Panel Study (NEPS) starting cohort 4 which include information on month of entry into child care. Schlotter (2011) uses data from the Children's Panel of the German Youth Institute (DJI-Kinderpanel).

where Y_{si} is the outcome s for child i; α_s is a constant; T_i is a treatment dummy; X_i is a vector of control variables that includes information on the child, his or her parents, as well as the household and the geographic region the child is living in. The parameter β_s will be estimated separately for each outcome variable; and ε_{si} is an error term.

However, even when conditioning on a rich set of covariates, estimating the above model by OLS will likely yield biased coefficients because of unobserved factors that influence both the treatment status and the outcome. For instance, parents may decide to send their children to child care earlier or later depending on their preferences for education. These preferences are likely not entirely captured by the parents' observable level of education and may be correlated with the level of support they are giving to their children in educational matters at home. Hence, if parents send their children to child care earlier and provide them with additional support, educational outcomes are favorably affected in two ways that cannot empirically be disentangled and will lead to inflated OLS estimates. One can also think of unobserved factors relating to the child in question that will bias "naïve" OLS estimates. One such factor is intelligence. Very intelligent children may be sent to child care earlier, because they are deemed ready by their parents earlier. At the same time, they will likely perform better on various skill and education measures. This notwithstanding, one can also imagine the opposite case to happen, namely less intelligent children to be sent to child care earlier in order to help them catch up with their peers. In that case OLS estimates will be biased downward. Further factors that may bias OLS estimates are spillover effects between siblings, for instance if a younger sibling benefits from child care attendance by an older sibling before entering child care him- or herself.

To mitigate these concerns, I apply an instrumental variables approach that relies on spatial variation in child care supply. Specifically, I instrument actual child care attendance at age 3 with the child care slot-child-ratio of 3- to 6.5-year-old children at the county level. This ratio is defined as the number of available child care slots for 3- to 6.5-year-old children divided by the number of children in this age group in the county where the child resides at age 3. It is calculated based on administrative data for the years 1994, 1998 and 2002 made available from the Federal Statistical Office of Germany (Statistisches Bundesamt, 2013). A detailed description of the variable is provided in section 2.5.2. Similar strategies have been used by Bach et al. (2019), Bauernschuster and Schlotter (2015), Cornelissen et al. (2018) as well as Felfe and Lalive (2018).[16] The idea behind the chosen approach is straightforward: In counties with higher slot-child-ratios, demand rationing will be less severe conditional on the set of control variables. Therefore, it will be easier for children in these counties to enter child care earlier – in this case at the age of 3 – than elsewhere. Via two-stage least squares regressions (2SLS) the local average treatment effect (LATE) of an additional year of center-based child care attendance on schooling outcomes for the group of *compliers* will be identified.[17] In the present case *compliers* are those children that enter into treatment solely because child care supply is higher in their county than elsewhere. Had they lived in a different county with lower child care supply, they would have entered child care one year later despite their parents' wish for them to enter at the age of 3.[18] In this case, the institutional surrounding makes for a natural experiment in which 3-year-old children are sorted in and out of child care at random.

The relationship between child care supply and actual attendance can be formalized as follows and serves as the first stage of the 2SLS IV estimations:

$$T_{si} = \pi_s + \varphi_s Z_i + \iota_s X_i + \mu_{si} \tag{2.2}$$

Here, Z_i is a continuous variable indicating the level of child care supply in a child's county of residence at age 3. Note that the vector X_i contains the same variables as described below Equation (2.1). The second stage model uses the results of the first stage to estimate the effect on schooling outcomes in adolescence. It looks as follows:

[16] Bauchmüller et al. (2014) use county-level information to instrument child care quality in Denmark. Next to higher level aggregates of the variables that they are interested in they also employ measures of regional political majorities, demographic changes as well as a policy governing (guaranteed) child care entry. The same policy is also used by Datta Gupta and Simonsen (2010) to instrument child care attendance.

[17] Strictly speaking, as most of the papers in the field I am estimating the combined effect of an additional year of child care attendance and a lower age at entry. The two effects cannot be disentangled with the data at hand.

[18] For a discussion of the informative value of the "compliant subpopulation", see Angrist and Pischke (2009, pp. 158).

$$Y_{si} = \kappa_s + \theta_s \hat{T}_{si} + \lambda_s X_i + \nu_{si} \tag{2.3}$$

In this model, $\hat{T}_{si}$ are the fitted values of child care attendance obtained from Equation (2.2) and θ_s describes the LATE estimate of the effect of one additional year of child care attendance on the various schooling outcomes.

In order for the coefficients θ_s to be consistent estimates of the effect of child care attendance on schooling outcomes, the slot-child-ratio has to not only be highly correlated with actual child care attendance while conditioning on a set of covariates (instrument relevance) but it also has to be uncorrelated with the outcome measures conditional on child care attendance and the various control variables (exclusion restriction). While instrument relevance can easily be tested in the first stage regressions, I support conditional independence of Z_i by conditioning on key predictors of county-level child care demand as well as on the mothers' personality as measured by the so-called Big Five.[19] The county-level demand indicators include the employment rate, the unemployment rate, log GDP per capita, the share of foreigners in the population, and the population density. As key predictors of local child care demand these indicators play a role in the administrative planning process of child care supply and must therefore not be omitted. The reason for including a measure of mothers' personality is that personality traits are likely correlated with schooling outcomes and may pick up hard-to-measure preferences that predict individual child care attendance while also potentially influencing the local institutional environment (i.e. the level of child care supply). In this context, one may be concerned about other unmeasured preferences for child care that are at the same time related to schooling outcomes and could therefore threaten the identification strategy. It is not possible to rule out this concern entirely. However, it is reassuring that a second widely used concept of a person's personality (and one that is not used in the main analysis), the locus of control as derived from seminal work by Rotter (1966), is orthogonal to local child care supply conditional on the vector of con-

[19] The "Big Five" are based on the five factor model of personality traits as developed by McCrae and Costa (1996). They include the concepts of conscientiousness, extroversion, agreeableness, openness as well as neuroticism (see e.g. John et al., 2008).

trol variables.[20] In this context one may also be concerned about non-random sorting of parents into counties with differing levels of child care supply. One obvious worry would be endogenous mobility, i.e. families with specific measured or unmeasured characteristics that have a bearing on schooling outcomes moving into areas with higher child care supply. However, previous research has found that inter-county mobility in Germany is low around childbirth and that migration patterns are not driven by child care supply (Felfe and Lalive, 2013; Cornelissen et al., 2018). Reassuringly, this conclusion is backed up by the data: Only 7 percent of all families moved to a different county in the first three years of their children's lives. More importantly, the incidence of inter-county migration is unrelated to child care supply rates.[21]

Given all of the above, I consider the remaining variation in child care supply to be plausibly exogenous. It is created by idiosyncratic administrative challenges and constraints faced by different counties, which are particularly pronounced in a rapidly changing environment as in the mid-1990s. Fittingly, Felfe and Lalive (2013) point out that the actual number of slots available to any given cohort is never identical to child care demand as a result of the lengthy administrative process in providing child care that bears the potential for planning errors, delays in approval and construction and shortages in the availability of pedagogical staff among others.

2.5 Data

2.5.1 The German Socio-Economic Panel Study

All individual-level data are taken from the German Socio-Economic Panel Study (SOEP). The SOEP is an annual representative household survey that was established in 1984 and has since been significantly expanded.[22] As of 2018 there are nearly 15,000 participating

[20] A mother's locus of control is measured by two variables that are constructed by factor analysis, a measure of the internal locus of control and a measure of the external locus of control. In both cases, the t-statistics of local child care supply range from 0.23 to 0.73 in absolute value (both when including and excluding the mothers' Big Five scores in the vector of control variables).

[21] Regressing a dummy variable whether or not the family has moved to a different country in the first three years of their child's life on regional child care supply (the instrument) yields a small positive coefficient of .017 (standard error .048) that is far from reaching statistical significance.

[22] For an overview, see Wagner et al. (2007).

households (Goebel et al., 2018). The longitudinal nature of the dataset allows following individuals over their life-course. The so-called Youth Questionnaire, which was introduced in 2000, provides information on a number of schooling outcomes when respondents are 16 or 17 years old. Answering the Youth Questionnaire marks the first time individuals from SOEP households are directly interviewed. All prior information is given by their parents and can easily be linked to the children.[23]

I only use information on adolescents, who have answered the Youth Questionnaire, who entered child care either at the age of 3 or at the age of 4 and whose families entered the SOEP prior to the child's third birthday. The latter restriction is necessary in order to obtain information on child care attendance in the year prior to legal eligibility, i.e. at the age of 2. That way, I am able to see if any child care attendance at the age of 3 or 4 marks the first time children are exposed to child care or if attendance is a mere continuation of previous attendance. In the latter case, observations are discarded as they would contaminate the treatment, which is defined as exactly one additional year of child care.[24] I further restrict the sample to children born in West Germany. The reason for this is that demand rationing was not an issue in the East as the former GDR provided extensive child care facilities free of charge to its citizens, many of which continued to exist after German reunification in 1990. Pooling across all available waves (1984 to 2016), 990 children who were born between 1982 and 1999 are left in the sample. To gain precision, I keep children born in the 1980s despite the fact that they were arguably not affected by the policy reform in 1992 that established the legal right to a place in child care for 3-year-olds. Nevertheless, demand rationing had been an issue for a while before and ultimately led to the passing of the law (Cornelissen et al., 2018). In the robustness section, I restrict the sample to those

[23] For more information on biographical data in the SOEP, see Frick and Goebel (2011).

[24] Strictly speaking, children could still have attended child care when they were 0 or 1 years old. However, very few children in Germany did so at the time of the survey. In my final sample, only 20 out of 1,075 children (1.86 percent) for whom there is information visited some form of center-based child care at the age of 1. The share is even lower at 0.58 percent among children aged 0. Importantly and as a result of my sample restriction, these children all have breaks in their child care careers later on. It is therefore unlikely that any later child care attendance is a continuation of previous attendance. What is more, child care for children under 3 was very different from kindergarten in the 1980s and 1990s and mostly provided in different locations. I therefore condition on a gap year before actual attendance at 3 or 4 instead of no previous child care attendance in order to not lose too large a number of observations. Note here that observations would not only be lost because of previous attendance but also because many parents entered the sample after their child's birth and information on previous child care is missing.

birth cohorts most directly affected by the reform and the rapid expansion of child care slots that went hand in hand with it and show that the results are not markedly altered.

2.5.2 Treatment

The main treatment indicator is a binary dummy that takes on the value 1 if the child attended any form of center-based child care in the year following the start of the first school year after his or her third birthday *and* had not done so in the year before.[25] It takes on the value 0 if the child entered center-based care one year later. Of the 990 observations, 648 are classified as treated, i.e. entered child care at the age of 3, and 342 are classified as untreated, i.e. entered child care at the age of 4.[26]

Information on child care attendance is provided by the head of the household during the annual interviews that generally take place in spring. Since I am not able to observe the exact month of child care entry, the treatment indicator is in fact a proxy of actual child care attendance. Following other researchers in the field (Schlotter, 2011), I assume that every year-child-observation of child care attendance actually reflects one entire year of attendance.[27] While this should be true for the majority of cases since entry into child care was in most places streamlined with the start of the school year, exceptions who entered child care during the school year are certainly present. These exceptions cause measurement error in the main independent variable, which will result in attenuation bias in OLS estimations and

[25] The start of the school year varies by year and by the 16 German states. With the data at hand, it is impossible to precisely assign students to one school year or the other since there is no information on their exact day of birth. I set the start of the school year to the end of September when the school year has started in all states. In the robustness section I discover the effects of leaving out "marginal" students born in August or September and show that this has only small effects on the results.

[26] In their study of the same policy reform, Bach et al. (2019) define the treatment as entering child care in the year of a child's third birthday not at the start of the next school year after the third birthday. This definition leads to lower attendance levels, as children born in the second half of the year are only classified as treated if they entered child care prior to their third birthday. I abstain from using this definition as it is not in line with the official cut-off rules and because the first stage relationship between regional child care supply and the treatment is weaker with this definition leading to larger standard errors.

[27] In his study of child care attendance on secondary track choice of 10- to 11-year-olds, Schlotter (2011) uses the same SOEP dataset and likewise assumes that every year-child-observation of child care attendance reflects one year of actual child care attendance.

bias the true effect of child care on schooling outcomes towards zero. However, this problem can be mitigated in the 2SLS models.[28]

2.5.3 Outcomes

This chapter deals with medium- and long-run effects of child care on a variety of schooling outcomes. As such, all outcome variables are taken from the Youth Questionnaire. The outcomes can be broadly grouped into three dimensions: those related to trajectories through the German school system, those related to cognitive skills, and those related to further educational aspirations. In the first group, I am interested in secondary track choice and grade repetitions. Track choice is measured via two binary dummies. The first one indicates whether or not a student attends the academic track (*Gymnasium*) of the German school system as opposed to students attending any other track, students who have left general schooling with a lower-level degree, and students who have dropped out of the general school system.[29] The second dummy indicates whether an individual attends the lowest track of the German school system (*Hauptschule*) or has already obtained a degree from this type of school as opposed to attending a higher track or having already obtained a higher-level degree. Note that the lowest secondary degree is usually awarded after 9 years of general schooling. Using different margins seems sensible as previous research has found differential effects of child care attendance on different socio-economic groups who often visit different secondary school tracks in Germany. Note that students are usually tracked in the German school system at the age of 10 upon completion of 4 years of primary schooling. Grade repetitions are also measured via a binary dummy that expresses if an individual has ever repeated a grade.

[28] Recall that measurement error only leads to attenuation bias if it affects the independent variable. In 2SLS estimations, the main treatment variable, which is measured imperfectly, becomes the dependent variable of the first stage regression circumventing the problem of independent variable measurement error.

[29] There are three lower tracks (*Gesamtschule*, *Realschule*, *Hauptschule*) in the German school system. While *Gesamtschule* may lead to a certificate that allows students to take up tertiary education, the other two are mainly geared towards the take-up of vocational training.

Cognitive skills are measured via grades on the last transcript in the subjects of language and mathematics.[30] For each subject, I use three different variables: one measuring grades linearly, one measuring whether or not a student has obtained one of the two top grades and one measuring whether or not a student has received one of the three worst grades. Note that German students are graded on a scale from 1 to 6 with 1 reflecting the highest level of achievement while both 5 and 6 mean failure of a course.[31] Importantly, in the linear specifications grades are standardized within school tracks in the full sample. That way, systematic differences in grading in different school tracks are leveled out and I am able to include all general schools in the estimations. The results are best interpreted as illustrating the effects of longer child care attendance on the relative performance of students vis-a-vis other students of the same school track and therefore complement the analyses of the effects on track choice.

Finally, educational aspirations reflect the highest professional degree that students plan to obtain after they have finished general schooling.[32] Possible answers are (1) an academic tertiary degree, (2) a vocational training degree, or (3) no professional degree. Based on the responses I construct three binary variables. The first one divides the children into those who aspire to a professional degree (indicated by response options (1) or (2)) and those who do not aspire to a professional degree, i.e. those who opted for response option (3). The other two variables take a closer look at the different possible margins. First, one variable expresses whether or not a tertiary degree is planned versus all other possible

[30] The sample includes both students who still attend one of the four tracks of general schooling and students who attend a vocational school. Since interviews almost exclusively take place in spring, most of these students should have already obtained a transcript from their vocational school in winter upon completion of the first or third semester. This is important since otherwise (i.e. if the most recent grades were obtained during general schooling) grades would pertain to different school tracks and thus not be comparable. In any event, leaving out students from vocational schools does not qualitatively alter the results.

[31] Ideally, I would have liked to only look at fail grades, i.e. 5 and 6, versus all other grades instead of 4, 5, and 6 versus all other grades. However, due to the relatively low number of students who obtain these grades this was not possible with the data.

[32] It is important to distinguish between educational aspirations and educational expectations. While the former pertain to goals and plans, the latter also take into account the self-reported probability of reaching these goals. Some authors have questioned that aspirations can be used as a vehicle to improve academic achievement (if high aspirations are not aligned with high expectations). However, in a recent study that deals with several cases of misaligned aspirations and expectations Khattab (2015) shows that high aspirations have a positive effect on school achievement even if expectations are low as compared to students with low aspirations and low expectations.

Table 2.1: Outcome indicators by treatment status

	Mean		Mean difference
	Control group	Treatment group	
PANEL A: ***Educational trajectories***			
Highest track	.311	.363	-.052
N	331	628	
Lowest track	.250	.182	.068**
N	324	610	
Repeater	.232	.230	.002
N	340	647	
PANEL B: ***Cognitive skills***			
Language grade	.169	.141	.028
Top language grade	.249	.297	-.048
Bottom language grade	.249	.261	-.012
N	313	597	
Mathematics grade	.089	-.045	.135*
Top mathematics grade	.280	.378	-.098***
Bottom mathematics grade	.341	.324	.017
N	311	596	
PANEL C: ***Educational aspirations***			
Degree (yes/no)	.905	.929	-.023
N	338	646	
Tertiary degree	.407	.497	-.090**
N	307	595	
Vocational degree	.850	.869	-.019
N	213	350	

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports mean values of all outcome measures for treated and untreated individuals at the age of 16 or 17, i.e. of those who entered child care at the start of the first school year after their third birthday and those who entered one year later and as a result spent one year less in child care. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999), own calculations.

degrees while the second variable captures whether or not students aspire to vocational training as compared to no further degree.

Table 2.1 reports mean comparisons of the outcome variables between treated and untreated individuals as well as the number of observations in each group. It can be seen that treated children outperform untreated children on almost all considered outcome measures. The one exception is the incidence of obtaining a bad grade in German language (panel B, upper part). In 4 out of 12 cases, the difference in means between treatment and control group reaches statistical significance. This is the case in the likelihood of attending the lowest secondary school track (panel A, middle), the average grade in math (panel B, lower part), the share of those obtaining a top grade in math (panel B, lower part), and the aspirations towards a tertiary education degree (panel C, middle). The gap is most pronounced in the incidence of top mathematics grades: At 38 percent, the share of students reaching such a grade is 10 percentage points higher in the treatment group than in the control group (28 percent).

2.5.4 Controls

Control variables pertain to the child in question, his or her parents and home environment as well as the county he or she is living in. Child characteristics that are controlled for include gender, migration status as well as full sets of year of birth and month of birth fixed effects. Year of birth fixed effects are necessary to capture all secular changes in schooling outcomes that occur over the period under study of almost two decades and that may be correlated with child care attendance. By including month of birth fixed effects I can eliminate differences in the age at child care entry between treatment and control group except those stemming from entering at the age of 3 or 4, since only children born in the same month of the year are compared with one another.[33] To account for differences in treatment intensity,

[33] Strictly speaking, I can only eliminate differences in age at child care entry if all children enter child care at the same time of the year, in this case at the start of the school year. This is the case for the majority of children, but not for all. Some minor differences in age at child care entry therefore remain. These differences are likely highly correlated with differences in the duration of child care attendance since entering child care during the school year most likely leads to a violation of the assumption that all child-year-observations of child care attendance actually reflect a whole year of attendance. The two effects cannot be disentangled. As described in section 2.5.2. this will lead to some attenuation bias in the estimates, which, however, can be mitigated in the 2SLS models.

I also discriminate between half-day and full-day attendance.[34] Parental characteristics include mother's level of education (tertiary, upper secondary, lower secondary or less), mother's personality (Big Five standardized), mother's age at childbirth as well as mother's employment status (full-time, part-time, not employed) while household controls comprise log income, total number of children dummies, birth order dummies, and an indicator on whether or not the child is brought up in a single parent household.[35] County-level controls comprise the employment rate, the unemployment rate, the share of foreigners in the population, the population density as well as log GDP per capita as described in section 2.4. All time-dependent control variables are measured three years after childbirth, i.e. at the time when treated children first become eligible to enter child care. In order not to lose any observations I impute missing data on all control variables. This is done by adding a missing category for dummy variables and by assigning the observation the respective mean value plus adding a missing dummy for continuous variables.

Table 2.2 shows mean values of control variables for treated and untreated individuals. There is some evidence of systematic selection into longer child care. At the individual level, having a migration background is negatively related to being in the treatment group. Furthermore, having a mother with a tertiary education degree exhibits a positive relation to being in the treatment group while the opposite is true for having a mother with no more than lower secondary education. These relationships are expected and corroborate previous research findings on a social gradient in child care take-up in Germany (Cornelissen et al., 2018; Felfe and Lalive, 2018; Jessen et al., 2019; Schober and Spieß, 2013; Scholz et al., 2019). In the lower part of panel B, it can be seen that living in a very large household of 4 or more children is negatively related to longer child care attendance. I speculate that in a lot of such very large families one parent decides to stay home full-time thereby obviating the need for center-based child care. Turning to the county-level controls, I observe signifi-

[34] Children are sorted into half- or full-day attendance according to the treatment regime that is observed more frequently before they enter primary school In the case of an equal number of observations, children obtain full-day status.

[35] If there was no information on the mother, information on the father was taken as a replacement. This is mostly the case for children whose parents are separated and who live in the household of their father.

Table 2.2: Balancing of sample on control variables

	Mean			Mean difference	*N*
	All	Control group	Treatment group		
	(1)	(2)	(3)	(4)	(5)
PANEL A: ***Individual controls***					
Gender	.51	.53	.51	.02	990
Migration background	.31	.34	.29	.05*	990
Full-day care	.19	.18	.21	-.03	990
PANEL B: ***Family-related controls***					
Maternal education					
Tertiary	.08	.06	.10	-.04**	986
Upper Secondary	.66	.64	.68	-.04	986
Lower Secondary or less	.26	.30	.23	.08***	986
Age at childbirth	28.4	28.2	28.6	-0.4	990
Maternal employment	.34	.32	.35	-.03	
Full-time	.08	.07	.08	-.01	983
Part-time	.26	.24	.27	-.03	983
No employment	.66	.68	.65	.03	983
Mother openness	.05	.04	.05	-.01	948
Mother conscientiousness	-.04	.00	-.06	.06	948
Mother extroversion	-.00	-.03	.01	-.04	948
Mother agreeableness	-.00	.03	-.02	.05	948
Mother neuroticism	.01	.03	-.01	.04	948
Household income	10.12	10.06	10.16	-.10	988
No. of children in HH					
1	.30	.28	.30	-.02	988
2	.48	.45	.50	-.04	988
3	.15	.15	.15	.00	988
4 or more	.07	.11	.06	.06***	988
Birth order					
1	.37	.39	.37	.02	887
2	.40	.37	.42	-.05	887
3	.15	.15	.16	-.01	887
4 or higher	.07	.08	.06	.03	887
Single parent	.03	.04	.03	.00	989
PANEL C: ***County-level controls***					
Employment rate	48.8	48.4	49.1	-.65***	983
Unemployment rate	8.77	9.11	8.59	.51**	983
GDP	3.19	3.18	3.20	-.02	969
Share of foreigners	10.1	9.8	10.3	-.51	988
Population density	535.4	473.1	568.4	-95.3	990

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports mean values and number of observations of control variables for the whole sample as well as for treated and untreated individuals at the age of 16 or 17, i.e. of those who entered child care at the start of the first school year after their third birthday and those who entered one year later and as a result spent one year less in child care. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999), own calculations.

cant mean differences between treated and untreated children on two of the five measures, namely the employment rate (positively related to longer attendance) and the unemployment rate (negatively related). This is not surprising as parental employment is naturally a major reason for demanding child care. Finally, in column 5 of Table 2.2, it can be seen that the number of missing observations on control variables is rather small and only in one case (birth order) larger than 5 percent.

When interpreting Table 2.2 it is important to bear in mind that we are dealing with a situation of undersupply of child care slots. Thus, differences between treated and untreated individuals are not entirely the result of intentional sorting but are also influenced by a degree of chance based on whether or not a slot was available when the child became eligible. This latter effect likely masks some of the pure preference-related differences and may explain why no find significant relationships are found on some measures where they could be expected such as in the case of maternal employment. In any event, in the framework of Altonji et al. (2005) the fact that there is some selection of children into longer child care on observable characteristics suggests that there may equally be selection on unobservable characteristics. This underscores the need for a quasi-experimental identification strategy such as the two-step IV approach outlined above.

2.5.5 Administrative Records

Administrative data on regional child care coverage rates for 3- to 6.5-year-old children are obtained from the Deutsche Jugendinstitut (DJI) and are based on records of the Federal Statistical Office of Germany (Statistisches Bundesamt, 2013). Information for all counties is available for the years 1994, 1998, and 2002. Children are assigned the respective rate of the year most closely coinciding with their third birthday. For example, children born in 1992 who turn 3 in 1995 are assigned the rate for 1994 while children born in 1994 are assigned the rate for 1998. Those children who turn 3 in the middle of two four-year-periods, for instance in 1996, are assigned the rate for the first of the two years. The resulting variable of child care supply has a mean of .78 and a standard deviation of .14.

2.6 Results

2.6.1 First Stage Regressions

The results of the first stage regression provide a hint at the strength of regional child care supply as an instrument for actual child care attendance. Theoretically, a higher regional child care supply should lead to more children being treated than elsewhere in a situation of demand rationing when everything else is equal as parents find it easier to secure slots for their children. To see that this is indeed the case, consider Table 2.3, which reports the first stage results of the 2SLS regressions for all considered outcome variables. Note that there is more than one first stage since I use different samples for the different outcomes. It quickly becomes obvious that the regression results reported in columns 1 through 8 are relatively similar. This is not surprising since the majority of observations are the same in all samples. I estimate that a one percentage point increase in the slot-child-ratio of 3- to 6.5-year-olds at the county level increases the probability of being treated (i.e. entering child care at the start of the first school year after ones third birthday) by between .905 (repeater, column 3) and .967 (low track, column 2) percentage points, thereby almost exactly yielding a one-to-one relationship. All estimates are highly significant. The strength of the instrument is underscored by the various first stage F-test statistics of between 24.0 (vocational degree, column 8) and 42.9 (low track, column 2).

2.6.2 Main Results

Table 2.4 presents the main estimation results of one additional year of child care attendance on schooling outcomes. Columns 1 to 3 provide OLS results that, however, may suffer from omitted variable bias. Columns 4 to 6 show 2SLS IV results that correct for such bias and show the average causal effect of an additional year of child care attendance for the group of compliers who expand their attendance as a result of more available slots in their county of residence. In all specifications, standard errors are clustered at the county level. Control variables are added to the models from left to right.

Table 2.3: First stage regression results

	First stage IV estimation							
	Educational trajectories			Cognitive skills		Educational aspirations		
	High track	Low track	Repeater	German language	Math	Some degree	Tertiary degree	Vocational degree
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
County-level childcare supply	.931*** (.134)	.967*** (.137)	.905*** (.136)	.937*** (.139)	.944*** (.139)	.917*** (.136)	.941*** (.138)	.959*** (.185)
Full controls	yes	yes	yes	yes	yes	yes	yes	yes
First stage F-test	38.9	42.9	35.9	38.6	39.3	37.1	36.3	24.0
N	959	934	987	910	907	984	902	563
R^2	0.217	0.218	0.210	0.211	0.208	0.210	0.221	0.254

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports first stage results of 2SLS IV regressions of one additional year of child care attendance on schooling outcomes. The child care slot-child-ratio of 3- to 6.5-year-olds at the county level is used as an instrument for actual child care attendance. Standard errors are clustered at the county level and given in parentheses. Individual control variable include gender, migration status, year of birth fixed effects, month of birth fixed effects, treatment intensity, mother's level of education, mother's age at childbirth, mother's employment status, mother's personality (Big Five), household income, number of children in household dummies, birth order dummies, and a single parent dummy. At the county level, control variables include employment, unemployment, GDP, share of foreigners, and population density. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999) and Statistisches Bundesamt (2013), own calculations.

Panel A of Table 2.4 shows the effects of longer child care attendance on educational trajectories. The baseline OLS estimates with just individual control variables presented in column 1 suggest that longer child care attendance may increase the probability of attending the highest school track at age 17 and decrease the probability of attending the lowest track. However, the coefficients considerably shrink in size and lose statistical significance when family-related control variables are added in column 2. This implies positive sorting of children from more advantaged families into longer child care attendance and at the same time higher school tracks. The OLS results of no significant effects on track choice and grade repetitions are confirmed in the IV models. Column 6 provides the results from my

Table 2.4: Effect of one additional year of child care attendance on schooling outcomes

	OLS			2SLS			*N*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PANEL A: ***Educational trajectories***							
Highest track	.060* (.034)	.013 (.034)	.014 (.034)	-.151 (.194)	-.058 (.157)	.032 (.158)	959
Lowest track	-.061** (.031)	-.031 (.028)	-.034 (.028)	.324* (.168)	.276** (.132)	.182 (.112)	934
Repeater	.002 (.028)	.004 (.030)	.001 (.031)	-.203 (.147)	-.228 (.147)	-.143 (.133)	987
PANEL B: ***Cognitive skills***							
Language grade	.011 (.071)	-.015 (.071)	-.010 (.073)	-1.12** (.466)	-1.11** (.434)	-.789** (.383)	910
Top language grade	.019 (.034)	.024 (.034)	.023 (.034)	.327 (.201)	.352* (.192)	.195 (.172)	910
Bottom language grade	.009 (.030)	.005 (.030)	.006 (.031)	-.410** (.166)	-.420** (.163)	-.321** (.154)	910
Math grade	-.068 (.072)	-.078 (.071)	-.083 (.072)	-.590 (.412)	-.574 (.388)	-.564 (.352)	907
Top math grade	.071** (.034)	.068* (.035)	.068* (.035)	.304* (.175)	.332* (.173)	.249 (.160)	907
Bottom math grade	-.002 (.032)	-.001 (.032)	-.008 (.032)	-.127 (.183)	-.108 (.166)	-.155 (.158)	907
PANEL C: ***Educational aspirations***							
Some degree	.016 (.021)	.014 (.021)	.012 (.021)	.203** (.094)	.204** (.091)	.162* (.088)	984
Tertiary degree	.081** (.038)	.051 (.037)	.052 (.038)	-.285 (.217)	-.229 (.190)	-.036 (.183)	902
Vocational degree	.020 (.031)	.028 (.033)	.019 (.033)	.301** (.134)	.288** (.120)	.221 (.137)	563
Individual controls	yes	yes	yes	yes	yes	yes	
Family controls		yes	yes		yes	yes	
County controls			yes			yes	

Notes: * p < 0.10; ** p < 0.05; *** p < 0.01. This table reports results of OLS and 2SLS IV regressions of one additional year of child care attendance on schooling outcomes. The child care slot-child-ratio of 3- to 6.5-year-olds at the county level is used as an instrument for actual child care attendance. Standard errors are clustered at the county level and given in parentheses. Individual control variables include gender, migration status, year of birth fixed effects, month of birth fixed effects, and treatment intensity. Family controls include mother's level of education, mother's age at childbirth, mother's employment status, mother's personality (Big Five), household income, number of children in household, birth order, and a single parent dummy. County controls include employment, unemployment, GDP, share of foreigners, and population density. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999); Statistisches Bundesamt (2013), own calculations.

preferred specification that includes the full set of control variables. While the likelihood of attending the highest and the lowest track are slightly increased and that of ever having repeated a grade slightly decreased, the coefficients do not reach statistical significance at conventional levels. The finding of no significant effects on track choice corroborates previous research findings by Kühnle and Oberfichtner (2017) and Müller et al. (2013) in a different sample and empirical setup. However, it stands in contrast with evidence from France provided by Dumas and Lefranc (2012) who find a reduced likelihood of grade repetitions among other things. A possible explanation for this is the very different pedagogical approach of publicly funded child care in France as compared to Germany with a starker emphasis on structured forms of learning.

On a general note, it is obvious that in columns 4 to 6 both the coefficients and the standard errors from the 2SLS models are larger than their OLS counterparts. This is a common finding in the literature and can have different reasons.[36] First of all, in IV regressions we are dealing with a very specific population group, the compliers in the language of Angrist and Pischke (2009). Since the 2SLS estimates are solely based on information for these compliers, the variation decreases thereby rendering the estimated effects less precise and blowing up confidence intervals. Second, the group of compliers may be altogether differently affected by child care than the group under study in the OLS regressions. If that is the case, the estimated LATE from the 2SLS models will not be equal to the average treatment effect (ATE) which is measured by OLS. Third, in section 2.5.1.1 I argued that the OLS estimates likely suffer from attenuation bias due to measurement error. The 2SLS procedure mitigates this problem by using actual child care attendance as dependent variable (of the first stage regressions) rather than as independent variable. As a result, coefficient estimates will necessarily increase in absolute value. Finally, the fact that the inclusion of county-level covariates in column 6 as compared to column 5 does have an effect on the estimates may signify that other unobserved county-level variables could have similar effects. However, I do not expect unobserved factors to influence estimates to a similar degree since the chosen county-level controls are key determinants of child care supply and in some part the very variables that are used by social planners to pick the appropriate level of supply. This notwithstanding, if relevant county-level variables remain omitted, the estimates would

[36] For instance, in Felfe and Lalive (2018) as well as Bach et al. (2019) similar patterns can be observed.

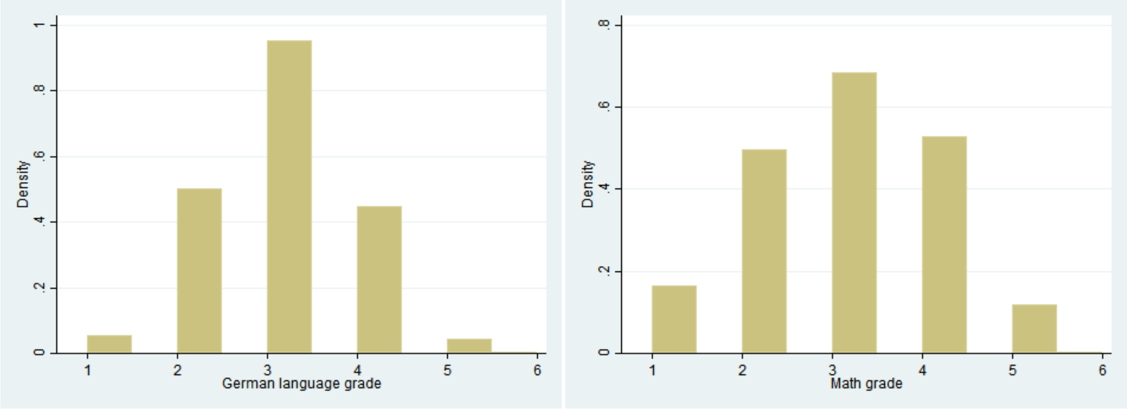
have to be interpreted as upper bounds on the true effect of longer child care attendance since most coefficients are depressed towards zero due to the inclusion of county-level controls throughout panels A to C.[37]

More significant effects than in the case of educational trajectories are found for cognitive skills in panel B of Table 2.4. The preferred IV estimates in column 6 show that German language skills are positively affected by an additional year of child care. In the linear specifications I obtain a large negative coefficient which suggests that longer child care attendance leads to grade improvements of 79 percent of a standard deviation. This estimate seems quite large even if it is an upper bound on the true effect. Indeed the confidence interval is rather wide. In any event, the estimate assumes linearity, which does not have to be the case. To get a better idea of who potentially profits from child care attendance I also look at two important margins in the grading system. The first is the likelihood of obtaining at least a 2 (the second-best grade) and the second is the likelihood of obtaining a 4 or worse. A look at Figure 2.1, which plots the density distributions of all German language and math grades in the estimation sample underscores the relevance of these margins. A very large portion of all students receives the "middle" grade 3 indicating a "satisfactory" level of achievement. This is true for 46 percent of all students in German language and 32 percent of all students in math. The grades of 2 and 4, which reflect a "good" and a "sufficient" level of achievement, respectively, are the next-most frequently obtained grades and therefore seemingly attainable for a large portion of students. In fact, the remaining three grades together account for only slightly more than 5 percent of all grades in German language and around 15 percent in math.

The results of the non-linear grade specifications are found in the rows named "top language grade" and "bottom language grade" as well as "top math grade" and "bottom math grade" in panel B of Table 2.4. They indicate that both considered margins are favorably affected by longer child care attendance. At 19.5 and 24.9 percent, the probability of obtain-

[37] Against this background, one should not be worried about a potentially detrimental effect of child care on track choice at the lower end of the track spectre as suggested by columns 4 and 5 of Table 2.4. After all, the effect becomes considerably smaller upon inclusion of county-level covariates in column 6 and loses statistical significance.

Figure 2.1: Density distribution of German language and math grades in the estimation sample



Source: SOEP v33 (1984-2016, birth cohorts 1982-1999).

ing a top grade (i.e. at least a 2) is increased in a very similar fashion in German language and mathematics. However, statistical significance is not given in the preferred specifications in column 6. Still, the math estimate comes rather close. The effect on the probability of obtaining a 4 or worse is less uniform in the two subjects. While the probability is significantly reduced in German language by 32.1 percent, the coefficient in math is half this size and far from reaching significance.[38] Again, the IV results are considerably larger than the OLS results. In fact, only the probability of obtaining a top math grade is significantly positively affected in the OLS models. Taken together, the results presented in panel B suggest that longer child care attendance may have positive causal effects on cognitive skills particularly among weaker students in German language. This finding stands in contrast to Kühnle and Oberfichtner (2017) who find no significant effects. However, there are a number of potential explanations that may reconcile our findings with theirs: First, by studying children who enter child care *before* they become eligible, Kühnle and Oberfichtner (2017) study a very selective group of people that may be less affected by public interventions due to high quality alternative modes of care. Second, significant effects are less likely to be found in the RDD design used by Kühnle and Oberfichtner (2017) since their treatment is

[38] The interpretation of these findings is facilitated by the fact that no effect of longer child care attendance on tracking is found, which could have introduced downward bias. The latter would have occurred because some treated children would have been shifted to the next higher track. It is easy to see that this will cause downward bias in each track as long as these marginal children perform worse on average than their peers in their new track.

much shorter in duration (four months as compared to a whole year in the present study). And finally, Kühnle and Oberfichtner (2017) are studying a different cognitive skill measure, namely test scores. In fact, in other countries significant positive long-run effects of child care on cognitive skills in general and school grades in particular have been reported (see e.g. Datta Gupta and Simonsen, 2016).

We now turn to panel C of Table 2.4 where the effects on educational aspirations are presented. Once more, the IV results are larger in size and show more significant effects that, however, mostly fade when controls are added. Only the probability of aspiring to a tertiary or vocational degree versus no degree at all is elevated by 16.2 percent due to an additional year of child care. When looking at the different margins in the lower two rows of panel C, it becomes obvious that this effect is driven by lower aspiring children who make a decision between a vocational degree and no degree at all. The likelihood of aspiring to a vocational degree is significantly increased in columns 4 and 5 and still positively affected, though not significantly so, in column 6. In contrast to this, the probability of aspiring to a tertiary degree does not seem to be affected at all by longer child care attendance as reflected by a coefficient that is very close to zero.

2.6.3 Sensitivity Analysis

The results of the previous section imply that longer child care attendance can have positive effects on cognitive skills and aspirations towards further education while essentially leaving track choice in secondary school and the probability to repeat a grade unaffected. In this section, I test the sensitivity of these results with regards to alterations of the sample under study as well as the delineation of the school year. The results of this exercise are presented in Table 2.5. In the first column, all children who have for at least one year been cared for by a childminder before entering school instead of attending center-based care are excluded from the analysis. In the main analysis, children are considered not treated if such care by a childminder happened in the year relevant for treatment, i.e. the year following their third birthday.[39] Therefore, eliminating these children should identify the pure effect of one addi-

[39] Considering children as treated if they have been cared for by a childminder in the year following their third birthday yields identical results to the baseline estimates up to the third decimal place. This procedure

tional year of center-based child care as opposed to family care instead of the effect of one additional year of center-based child care versus family care or care by a childminder.[40] Given the fact that the sample is only marginally reduced, it is not surprising that the results of the main analysis are basically confirmed.[41] However, it is noteworthy that all coefficients slightly move in the direction of greater benefits due to longer child care attendance. What is more, the previously insignificant coefficients on math grades and aspirations towards a vocational degree become weakly significant.

A second sensitivity check pertains to the birth cohorts under study. Arguably, children born in the early and mid-1980s were less affected by the child care reform in 1992 since the expansion of child care supply and the debate about a legal right to a place in child care for 3-year-olds only gathered steam at the beginning of the 1990s. I therefore restrict the sample to children who were born in 1987 or later and therefore turned 3 in 1990 or later. In total, 252 observations are discarded. First stage results of the 2SLS IV estimations using the remaining observations are presented in column 2 of Table A2.1 in the appendix. It can be seen that the strength of the instrument as measured by the F-test statistic is not significantly altered compared to the larger sample in the baseline estimations presented in Table 2.3 and is in three subsamples even superior to it. In fact, all first stage coefficients are larger than in Table 2.3, although standard errors are also larger due to the smaller sample size. Taken together, the first stage results lend support to the claim that children who started their child care careers in the 1990s were most affected by the reform. Turning to the second stage estimation results in column 2 of Table 2.5, we see that the coefficients on all outcomes are quite similar to those in column 6 of Table 2.4 but less precisely estimated. For this reason the effects on German language grades narrowly fail to reach statistical sig-

could be interpreted as giving an estimate of the effect of one additional year of all forms of away-from-home care.

[40] While children who have been cared for by a childminder in the year after their third birthday are considered not treated, some treated observations are also lost due to the procedure conducted here. The reason for this is that some treated children have experienced care by a childminder later in the child care careers. In that sense, excluding all children who have ever experienced such care should give a better proxy of the actual treatment than the one used in the baseline estimations. However, since information on childminder care is only available from 1995 onwards, some older children who have been cared for by a childminder may be left in the sample. The movement in coefficients resulting from excluding childminder children could therefore in reality be somewhat larger than indicated by the results in column 1 of Table 2.5.

[41] Consider Table A2.2 in the appendix for information on sample sizes of all sensitivity checks.

Table 2.5: Sensitivity of main results of the effect of one additional year of child care attendance on schooling outcomes

	2SLS		
	No childminder	Birth cohorts 1987+	No births in August or September
	(1)	(2)	(3)
PANEL A: ***Educational trajectories***			
Highest track	.061 (.161)	-.011 (.189)	.071 (.175)
Lowest track	.162 (.111)	.116 (.137)	.160 (.124)
Repeater	-.166 (.134)	-.168 (.140)	-.136 (.155)
PANEL B: ***Cognitive skills***			
Language grade	-.876** (.389)	-.694 (.426)	-.312 (.407)
Top language grade	.220 (.174)	.193 (.191)	.029 (.197)
Bottom language grade	-.350** (.155)	-.274 (.174)	-.168 (.169)
Math grade	-.717** (.362)	-.510 (.382)	-.695* (.421)
Top math grade	.322* (.165)	.125 (.172)	.356* (.204)
Bottom math grade	-.221 (.163)	-.224 (.167)	-.185 (.184)
PANEL C: ***Educational aspirations***			
Some degree	.167* (.089)	.168* (.092)	.153 (.104)
Tertiary degree	-.025 (.183)	-.119 (.195)	.040 (.215)
Vocational degree	.234* (.136)	.248* (.143)	.174 (.166)
Full controls	yes	yes	yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports results of 2SLS IV regressions of one additional year of child care attendance on schooling outcomes. The child care slot-child-ratio of 3- to 6.5-year-olds at the county level is used as an instrument for actual child care attendance. Standard errors are clustered at the county level and given in parentheses. Individual control variables include gender, migration status, year of birth fixed effects, month of birth fixed effects, and treatment intensity. Family controls include mother's level of education, mother's age at childbirth, mother's employment status, mother's personality (Big Five), household income, number of children in household, birth order, and a single parent dummy. County-level controls include employment, unemployment, GDP, share of foreigners, and population density. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999); Statistisches Bundesamt (2013), own calculations.

nificance.[42] On the contrary, the slightly increased coefficient on aspiring to a vocational degree as compared to no degree again turns out weakly significant. This further fortifies the conclusion that educational aspirations predominantly seem to be positively affected at the lower margin.

A final sensitivity check aims at shedding light on the role played by students born close to the cut-offs in August and September. It is not possible to precisely assign them to one or the other cohort since their exact date of birth is unknown and the school year is usually not aligned with the beginning of a new month. In the baseline estimations, I set the start of the school year to the end of September when the school year has started in all 16 German states. Some students may thereby be assigned to the wrong birth cohort. To investigate the extent to which these "marginal" students influence the results I exclude all children born in August and September from the regressions in column 3 of Table 2.5. Once more, most estimates are in the same range as before but statistical significance is lost due to higher standard errors. However, analogous to the models excluding childminder children now the estimates for math grades and particularly top math grades are somewhat increased so that they reach statistical significance at the 10-percent-level.

All in all, the results of the sensitivity checks confirm the baseline findings that attending child care for an additional year can have positive effects on cognitive skills and educational aspirations while educational trajectories are left unaffected. However, due to the even smaller sample sizes used in this section precision is compromised.

2.6.4 Heterogeneous Effects

In this section I check whether different groups of children are differently affected by longer child care attendance. I therefore split the sample along socio-economic status (SES) and gender lines. A low SES is assigned if a child's mother held a degree from the lowest educational track (*Hauptschule*) with no further vocational training or no degree when the child was 3 years old. Furthermore, I look into potential differences between children who were

[42] Statistical significance at the 10-percent-level can be reached by adding one or more adjacent cohort(s) to the ones used in this sensitivity check.

Table 2.6: Heterogeneity of first stage regression results

	First stage IV estimation					
	SES		Gender		Treatment intensity	
	High	Low	Boys	Girls	Half-day	Full-day
	(1)	(2)	(3)	(4)	(5)	(6)
County-level child-care supply	.847*** (.149)	1.370*** (.385)	.811*** (.202)	.959*** (.183)	.892*** (.154)	1.066*** (.380)
Full controls	yes	yes	yes	yes	yes	yes
First stage F-test	25.0	14.8	15.0	21.2	28.2	9.1
N	782	204	482	508	793	197
R^2	0.213	0.411	0.246	0.285	0.228	0.425

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports first stage results of 2SLS IV regressions of one additional year of child care attendance on schooling outcomes. The child care slot-child-ratio of 3- to 6.5-year-olds at the county level is used as an instrument for actual child care attendance. Standard errors are clustered at the county level and given in parentheses. Individual control variable include gender, migration status, year of birth fixed effects, month of birth fixed effects, treatment intensity, mother's level of education, mother's age at childbirth, mother's employment status, mother's personality (Big Five), household income, number of children in household dummies, birth order dummies, and a single parent dummy. At the county level, control variables include employment, unemployment, GDP, share of foreigners, and population density. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999); Statistisches Bundesamt (2013), own calculations.

predominantly cared for in half-day and full-day mode respectively. As a first step in this analysis it is important to examine whether all subgroups of children respond to the instrument. I therefore run separate first stage regressions for all subgroups using the full estimation sample. The results of this exercise are depicted in Table 2.6. They suggest that all subgroups are similarly affected by the instrument with coefficients lying in the range of .811 (boys, column 3) to 1.370 (low SES, column 2). Given the smaller sample sizes, this slight variability in coefficients should not be worrying.[43] However, in the case of full-day chil-

[43] If I were to make an attempt at interpreting the most striking differences in effect sizes, i.e. those between high and low SES children, one could argue that a smaller coefficient for high SES children makes sense

dren the instrument narrowly fails to pass the rule-of-thumb test of an F-statistic larger than 10 so the results for these children should be treated with some caution.

Table 2.7 presents 2SLS results that include interaction terms for the subgroups under study. The necessary second instrument is constructed by interacting regional child care supply with the subgroup dummy. In the first two columns the sample is split along SES lines. Column 1 reports the effect for high SES children and column 2 the interaction effect with low SES. For high SES children the results are very similar to the main results depicted in column 6 of Table 2.4. Save for a weakly significant negative interaction on top language grade in panel B, effects on low SES children generally do not seem to differ significantly from the baseline results, either. The significant interaction on top language grade should not be over-interpreted since the effect on high SES children is not significant and reverse coding reveals that the partial effect for low SES children is far from reaching statistical significance, too. The finding of no significant interactions may seem counter-intuitive at first glance, as most of the previous literature has found children from low socio-economic backgrounds to be particularly beneficially affected by center-based child care. One may hypothesize, however, that the mother's educational status is not a perfect indicator for singling out "problematic" families as many women in Germany did not pursue upper secondary or even tertiary education at the time anticipating that they would mainly care for their children later on. Arguably, looking at different margins in terms of aspirations and grades as is done above is therefore a more effective way of discovering social gradients in the effect of child care attendance. The main results depicted in Table 2.4 seem to confirm this as reflected by significant effects at the lower margins of school grades and educational aspirations.

Columns 3 and 4 report results for boys and the interactions for girls, respectively. In panel B, we see that any beneficial effect on grades is actually clustered among girls in both language and math. This is in line with a lot of previous research on both targeted interven-

since demand rationing tends to favor high SES children whose parents have more and better resources to secure one of the scarce child care places for their offspring.

Table 2.7: Heterogeneous effects of one additional year of child care attendance on schooling outcomes

	2SLS					
	SES		Gender		Treatment intensity	
	High		Boy		Half-day	
		x Low		x Girl		x Full-day
	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: ***Educational trajectories***						
Highest track	.050 (.189)	-.093 (.257)	-.153 (.224)	.328 (.291)	.074 (.167)	-.225 (.411)
Lowest track	.104 (.125)	.273 (.255)	.153 (.154)	.052 (.215)	.166 (.122)	.084 (.377)
Repeater	-.199 (.142)	.225 (.333)	-.113 (.166)	-.055 (.243)	-.145 (.139)	.014 (.437)
PANEL B: ***Cognitive skills***						
Language grade	-1.003** (.443)	.928 (.759)	-.007 (.527)	-1.371* (.700)	-.845* (.443)	.312 (1.562)
Top language grade	.308 (.199)	-.482* (.289)	-.122 (.215)	.556* (.294)	.231 (.200)	-.203 (.580)
Bottom language grade	-.355** (.173)	.180 (.323)	.008 (.237)	-.578** (.282)	-.318* (.169)	-.020 (.645)
Math grade	-.557 (.422)	.005 (.694)	-.040 (.497)	-.924 (.651)	-.282 (.422)	-1.581 (1.600)
Top math grade	.224 (.182)	.094 (.299)	.081 (.221)	.296 (.263)	.158 (.173)	.514 (.616)
Bottom math grade	-.211 (.188)	.252 (.346)	.051 (.218)	-.363 (.314)	-.069 (.193)	-.481 (.672)
PANEL C: ***Educational aspirations***						
Some degree	.144 (.098)	.124 (.181)	.326** (.128)	-.306* (.161)	.124 (.098)	.214 (.332)
Tertiary degree	.089 (.204)	-.451 (.404)	-.251 (.302)	.356 (.363)	-.016 (.205)	-.100 (.454)
Vocational degree	.192 (.147)	.138 (.197)	.418** (.192)	-.403* (.229)	.184 (.150)	.169 (.432)
Full controls	yes		yes		yes	

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports results of 2SLS IV regressions of one additional year of child care attendance on schooling outcomes. The child care slot-child-ratio of 3- to 6.5-year-olds at the county level is used as an instrument for actual child care attendance. Standard errors are clustered at the county level and given in parentheses. Individual control variables include gender, migration status, year of birth fixed effects, month of birth fixed effects, and treatment intensity. Family controls include mother's level of education, mother's age at childbirth, mother's employment status, mother's personality (Big Five), household income, number of children in household, birth order, and a single parent dummy. County controls include employment, unemployment, GDP, share of foreigners, and population density. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999); Statistisches Bundesamt (2013), own calculations.

tions and universal child care regimes that has found disproportionately larger cognitive gains from child care attendance for girls (see e.g. Anderson, 2008; Barnett et al., 1998; Campbell et al., 2002; Fessler and Schneebaum, 2016; Havnes and Mogstad, 2011; Heckman et al., 2013; Herbst, 2017; García et al., 2017; Sandner, 2013). On the contrary, educational aspirations seem to be more positively affected among boys as illustrated in panel C.

Finally, in columns 5 and 6 the sample is split by treatment intensity. For half-day treated children, who constitute a large majority of all children, there is only very little change in coefficients as compared to the main results. There are no significant interactions for full-day treated children, either. However, it has to be recalled that the first stage for this relatively small group of children was rather weak, which is reflected in very large standard errors that complicate meaningful interpretation of the interaction terms.

2.7 Conclusion

This paper presents evidence on the effects of an additional year of center-based child care attendance on a range of medium- and long-run educational outcomes for Germany. It exploits a policy reform in 1992 that granted the right to a place in publicly funded child care to all children aged 3 or older until the start of primary school. The ensuing staggered expansion of child care supply across counties is used for identification. The results suggest that especially weaker students may benefit from longer child care attendance. This is exemplified by a reduced likelihood of obtaining bad grades in German language and increased aspirations towards obtaining a vocational degree later on. No effects are found for secondary school track choice and the likelihood of grade retentions. These findings are generally in line with previous research and significantly expand the evidence base on the type of medium-term and long-term effects that can be expected from longer child care attendance.

When interpreting the results, one should bear in mind that the LATE effects estimated in this study pertain to children from families with a low resistance to child care. Since these are often relatively well-off and have high-quality alternative modes of care, one may hypothesize that other groups of children may be differentially and perhaps even more positively influenced by longer child care attendance. At the same time, I acknowledge that sample sizes in this paper are rather small, which adversely affects precision. Rather than pinpointing exact effect sizes the results of this study are therefore best understood as providing guidance on where to look for significant effects and what signs to expect. In this light, they prove that longer child care attendance should in no case be detrimental to child development even among early "takers" of center-based child care and may have positive medium- and long-run effects on cognitive skills as well as future educational aspirations. This further strengthens the case for public funding of child care centers, which also rests on the positive effects on maternal labor supply (for maternal employment effects of the same reform, see Bauernschuster and Schlotter, 2015). Especially policy-makers concerned with fighting inequality should find public involvement in child care a suitable means to reach their target, since, by benefiting particularly weaker students, longer child care attendance seems to be "leveling the playing field" not only in the short-run but also over a larger time horizon. However, in order to derive more definite policy implications, especially with regards to the net effects on the public purse, more research with larger sample sizes that will lead to more precisely estimated effect sizes is needed.

Appendix: Additional Tables

Table A2.1: First stage regression results for sensitivity analysis of main results of the effect of one additional year of child care attendance on schooling outcomes

	2SLS					
	No childminder		Birth cohorts 1987+		No births in August or September	
		F-test		F-test		F-test
	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: ***Educational trajectories***						
Highest track	.939*** (.136)	38.5	.940*** (.150)	38.0	.894*** (.154)	28.6
Lowest track	.975*** (.138)	42.3	.982*** (.153)	40.0	.933*** (.156)	31.9
Repeater	.912*** (.137)	35.6	.949*** (.149)	38.7	.869*** (.156)	25.9
PANEL B: ***Cognitive skills***						
Language grades	.942*** (.140)	37.8	.955*** (.156)	37.0	.906*** (.159)	28.3
Math grades	.950*** (.140)	38.6	.960*** (.155)	37.7	.918*** (.159)	28.9
PANEL C: ***Educational aspirations***						
Some degree	.924*** (.137)	36.9	.947*** (.149)	38.2	.887*** (.156)	27.6
Tertiary degree	.952*** (.139)	36.0	.976*** (.154)	37.3	.910*** (.157)	27.9
Vocational degree	.971*** (.187)	24.5	1.055*** (.207)	31.4	.926*** (.210)	18.5
Full controls	yes		yes		yes	

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports first stage regression results for 2SLS IV regressions of one additional year of child care attendance on schooling outcomes as conducted in Table 2.5. The child care slot-child-ratio of 3- to 6.5-year-olds at the county level is used as an instrument for actual child care attendance. Standard errors are clustered at the county level and given in parentheses. Individual control variables include gender, migration status, year of birth fixed effects, month of birth fixed effects, and treatment intensity. Family controls include mother's level of education, mother's age at childbirth, mother's employment status, mother's personality (Big Five), household income, number of children in household, birth order, and a single parent dummy. County-level controls include employment, unemployment, GDP, share of foreigners, and population density. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999) and Statistisches Bundesamt (2013), own calculations.

Table A2.2: Number of observations for sensitivity analysis of main results of the effect of one additional year of child care attendance on schooling outcomes

	2SLS		
	No childminder	Birth cohorts 1987+	No births in August or September
	(1)	(2)	(3)
PANEL A: ***Educational trajectories***			
Highest track	930	724	772
Lowest track	905	708	750
Repeater	958	735	797
PANEL B: ***Cognitive skills***			
Language grade	881	685	732
Top language grade	881	685	732
Bottom language grade	881	685	732
Math grade	878	685	730
Top math grade	878	685	730
Bottom math grade	878	685	730
PANEL C: ***Educational aspirations***			
Some degree	955	734	794
Tertiary degree	874	684	722
Vocational degree	555	395	450
Full controls	yes	yes	yes

Notes: This table reports the number of observations in the 2SLS IV regressions of one additional year of child care attendance on schooling outcomes as conducted in Table 2.5. The child care slot-child-ratio of 3- to 6.5-year-olds at the county level is used as an instrument for actual child care attendance. Individual control variables include gender, migration status, year of birth fixed effects, month of birth fixed effects, and treatment intensity. Family controls include mother's level of education, mother's age at childbirth, mother's employment status, mother's personality (Big Five), household income, number of children in household, birth order, and a single parent dummy. County-level controls include employment, unemployment, GDP, share of foreigners, and population density. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Source: SOEP v33 (1984-2016, birth cohorts 1982-1999) and Statistisches Bundesamt (2013), own calculations.

CHAPTER 3

ENGAGING TEACHING PRACTICES AND ACHIEVEMENT – A WITHIN-STUDENT APPROACH IN THREE SUBJECTS

3.1 Introduction

It is well-established that teachers are important inputs in educational production (Hanushek and Rivkin, 2010; Koedel et al., 2015; Wößmann, 2003b). Equally well-established is that the monetary gains from improved teaching in schools would be quite substantial (Chetty et al., 2014; Hanushek, 2011). However, in order to improve the quality of teaching one has to identify what makes teachers effective in conferring skills upon students. So far, economists have not been very good at this. Socio-economic teacher characteristics such as gender, teaching experience, age, and education alone cannot account for the huge achievement differences attributable to different instructors (see e.g. Lavy, 2015). Rather, there is a need to analyze what teachers do in classrooms and how they interact with their students (Schacter and Thum, 2004).

In recent years, the advent and expansion of large-scale assessment studies such as PISA, TIMSS, and PIRLS has enabled researchers to take a closer look at the effects of specific teaching practices. However, most studies have limited themselves to the dichotomy of "traditional" versus "modern" teaching (Bietenbeck, 2014; van Klaveren, 2011; Lavy, 2015; Schwerdt and Wuppermann, 2011). This paper goes one step further and assesses the impact of engaging teaching practices on student achievement. The intuition behind this is that more engaged students learn at a faster rate than less engaged students. The engaging teaching practices investigated in this chapter are the use of questioning in class, bringing interesting materials to the course, relating the course content to students' daily lives, giving praise and encouragement and summarizing the most important points of the lesson.

The concept of student engagement is well-known in the educational sciences and has been described as capturing the ‘in-the-moment cognitive interaction’ of the student with what is being taught (McLaughlin et al., 2005). A meta-analysis of teaching effectiveness studies by Seidel and Shavelson (2007) illustrates its relevance: They find that active student engagement as a generalizable, not subject-specific input in learning is weakly positively related to cognitive and modestly so to non-cognitive achievement measures (Seidel and Shavelson, 2007). Other literature reviews find additional evidence that more engaged students achieve better learning results (Fredricks et al., 2004). Educational achievement in turn is one central predictor of labor market outcomes later in life (see e.g. Wößmann, 2016). In addition, higher engagement can also lead to lower incidence of delinquency, aggression, and early school dropout (Fredricks et al., 2004; Hill and Werner, 2006).[44] Against this background, knowing how to engage students in classrooms would be of great social value. Such knowledge could be incorporated into teacher training and, if successfully applied in classrooms, could yield substantial monetary and non-monetary rewards for students themselves and for society as a whole. However, a necessary precondition for this is that student engagement can be altered by outside influences. It is reassuring that researchers have found engagement to be relatively malleable, among other things by school and classroom factors (see e.g. Fredricks et al., 2004).

The 2011 waves of the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) contain information on teachers’ uses of engaging teaching practices. TIMSS and PIRLS are internationally comparative assessments of fourth- and eighth-graders’ achievement in math, science and reading. TIMSS deals with the two subjects of mathematics and science, while PIRLS is dedicated to reading. Both studies are administered by the International Association for the Evaluation of Educational Achievement (IEA), an organization that has been carrying

[44] In the educational literature, engagement is often defined as multifaceted construct encompassing several, intertwined dimensions. These include the affective sphere (do children like school?), a behavioral component (participating in schooling activities, doing homework) as well as a cognitive part (e.g. motivation) (Perdue et al., 2009). In the present analysis, I will use the somewhat narrower definition of engagement given by McLaughlin et al. (2005), which is also the basis for the indicator as constructed in the TIMSS and PIRLS.

out international assessments of student achievement since 1959. Via the two studies, it aims at enabling participating countries to improve their educational policy. As well as achievement data, both studies collect extensive background information on student, family, and institutional factors. Choosing the 2011 waves of the two studies is beneficial in two ways: First, 2011 was the first occasion on which the two studies included the so-called 'Engaging Students in Learning Scale' (ESL scale), which serves as my main explanatory variable of interest. Second, 2011 is the only year so far in which TIMSS and PIRLS were sampled together – at least for students in fourth grade. This provides the unique opportunity of observing primary school students in three different subjects, namely math, science, and reading. In total, 34 countries and 3 benchmarking entities participated in the joint sampling of TIMSS and PIRLS.[45] The full TIMSS and PIRLS 2011 database contains information for 185,475 students, 171,098 parents, 14,258 teachers, and 6,469 school principals (Foy, 2013). For a variety of reasons, I only focus on Germany, the largest European country. First of all, there are only very few countries in which a significant share of primary school students is taught by different teachers in different subjects. This, however, is a precondition for my identification strategy. Secondly, these countries are predominantly Arab countries that may be very different from Western countries in terms of learning approaches and learning content. Since learning is context sensitive, I also refrain from pooling the available data from other countries, as educational systems differ and teaching practices that are beneficial in one country are not necessarily equally beneficial in another country.

The fact that I observe test scores in multiple subjects for each student allows me to use within-student estimation for identification. By doing so, I am able to control for individual time-invariant characteristics that also affect achievement such as underlying ability, lagged achievement, parental background, and school resources. A major advantage for the identification of the desired effect is the fact that I am using primary school students as units of analysis. Most of these students have no long history of different teachers in different subjects, a fact that makes it likely that any observed teacher effect can be attributed to the current instructor. What is more, selective admission and within-school sorting by ability should be less of a problem in primary school than in secondary school. A caveat of the

[45] Benchmarking entities are regional jurisdictions within countries that participate separately in the assessments.

chosen approach is the assumption that the effect of engaging teaching practices on achievement is the same in all three subjects. This is a common assumption in empirical research on the economics of education. Yet, in this paper I can test and relax it in two ways. First of all, I estimate models in which only two subjects are pooled at the same time. Second, I estimate correlated random effects models that explicitly model the potential correlation of the unobserved heterogeneity with the observed inputs and, therefore, allow me to obtain subject-specific coefficients (other studies that rely on correlated random effects models are Falck et al., 2015; Metzler and Wößmann, 2012; Piopiunik and Schlotter, 2012).

This article adds to the literature by being the first to empirically estimate the effect of engaging teaching practices on achievement outcomes. What is more, to my knowledge it is the first paper that employs different within-student between-subject identification methods in the framework of three different subjects. It thereby expands the relatively scarce economic literature on the effectiveness of specific teaching practices both content-wise and methodologically. The results indicate that the use of engaging teaching practices as measured by the ESL scale has no significant effect on achievement. However, I do find a modest positive effect for students from low socio-economic backgrounds. A one-standard-deviation-increase on the ESL scale raises test scores in math, science and reading by 4.6 percent of a standard deviation. The latter is equal to about three points on the achievement test. This finding suggests that more use of engaging teaching practices can yield societal gains in terms of greater equality of opportunity. This is important in Germany, where intergenerational educational mobility is generally found to be low (Heineck and Riphahn, 2009). The finding is also in line with some previous research on teaching practices that has found differential results by subgroup under study (Lavy, 2015). What is more, there seems to be negative sorting of students to teachers in Germany. This is reflected by the fact that the ‘naïve’ OLS results tend to be more negative than the fixed effects estimates. Such sorting could for instance happen if parents of comparably worse students send their children to schools with teachers who put particular emphasis on engaging students. It could also be the result of teachers adjusting their behavior to a class of low performing students in the sense that they try to more actively encourage their students to get involved in the subject matter. Overall, the results are robust to a variety of sensitivity checks. Still, the correlated random effects models estimated in the latter part of the paper suggest that subject differences in the estimated effects cannot entirely be ruled out. Specifically, trying to engage students seems to be more effective in reading than in other subjects. Generally, it should be noted that the within-student between-subject approach used in this paper does not allow me to entirely rule out bias stemming from unobserved teacher characteristics. This weakens the interpretation of my results as reflecting causal mechanisms.

The remainder of this chapter is structured as follows: Section 3.2 provides an overview of the relevant economic literature. Section 3.3 outlines the identification strategy. Section 3.4 introduces the data, the analysis sample and the central variables used in the estimations. Section 3.5 presents the results as well as several robustness checks. Finally, section 3.6 concludes.

3.2 Related Literature

A number of economic studies have attempted to gain insights into educational production by relating easily observable teacher characteristics such as age, gender, education, and experience to student achievement. The main result from these studies is that teaching experience has a positive effect on student achievement (Clotfelter et al., 2006; Goldhaber and Anthony, 2007; Rivkin et al., 2005). However, the effect appears to be non-linear, leveling off after around five years (Rivkin et al., 2005). Most other characteristics are generally found to have either negligible or no effects on achievement (Aaronson et al., 2007; Rivkin et al., 2005). Significant effects can mostly be found for certain subgroups of students or specific student-teacher pairings on certain characteristics. For instance, Paredes (2014) finds that teacher-student gender matching can have positive effects on performance via role model effects.

Since readily observable teacher characteristics can only explain a small fraction of the variation in student achievement, researchers have tried to go beyond analyzing objective traits and attempted to assess what happens in classrooms. The data for this are typically gathered in one of three ways: They are either based on (1) classroom observations by trained experts, (2) student reports, or (3) teacher self-reports. Prominent examples of the first group of studies are provided by Tyler et al. (2010) and Kane et al. (2011) who use data collected by the Cincinnati Public Schools' Teacher Evaluation System (TES). They find that observational quality measures are clearly related to achievement outcomes. In an analysis of different components of the overall TES score, Tyler et al. (2010) find that teachers who place more emphasis on the classroom environment instead of focusing on specific teaching practices can reap particularly large achievement gains among their students. Similarly and particularly relevant for this research, teachers who engage their students in questions and discussions are more effective than teachers who routinely focus on additional content. This result, however, is only valid for reading, not for mathematics. The result of questioning and discussion being particularly effective in reading is corroborated by Kane et al. (2011). In contrast to this, Blazar (2015) fails to find significant effects for different dimensions of two routinely used observational instruments, namely the Mathematical Quality of Instruction (MQI) and the Classroom Assessment Scoring System (CLASS).

Studies based on student or teacher reports generally make use of large-scale assessment data. Most of these papers deal with the question of whether 'traditional' teacher-centered teaching or 'modern' student-centered teaching is more effective in conferring skills upon students. The former is characterized by strong reliance on lecturing and direct instruction, while the latter shifts the emphasis onto group work. Using teacher self-reports, Schwerdt and Wuppermann (2011) find tentative evidence that traditional lecture-style teaching is superior to modern teaching. However, they admit that their results may be influenced by selection bias and conclude that traditional teaching is at least not worse than modern teaching. Van Klaveren (2011) finds no significant effect of lecture-style teaching, while other studies provide some evidence of explicitly negative effects of some elements of modern teaching (Goldhaber and Brewer, 1997; Hidalgo-Cabrillana and López-Mayan, 2015; Murnane and Phillips, 1981).

Postulating that modern and traditional teaching methods can coexist alongside each other, Lavy (2015) finds large payoffs for both traditional and some facets of modern teaching, which do, however, differ by subgroup under study. While girls and students from low socio-economic backgrounds seem to benefit most from teacher-centric education, students from higher socio-economic backgrounds can be especially well targeted by modern teaching methods. Using TIMSS data for US eighth-graders, Bietenbeck (2014) demonstrates that traditional and modern teaching methods promote different skills in children. While traditional teaching is particularly useful for increasing students' factual knowledge, modern teaching improves reasoning skills. Both Lavy (2015) and Bietenbeck (2014) rely on student reports to measure teaching styles.

There are few economics papers based on large-scale survey data that go beyond the dichotomy of modern and traditional teaching and try to shed light on specific teaching practices.[46] A notable exception is provided by Aslam and Kingdon (2011), who use teacher self-reports in Pakistan and find that certain teaching practices are significant predictors of student outcomes. This is especially true for the use of quizzing and questioning in class as well as planning of the lesson at home. However, their results are mostly confined to private schools and originate from a developing country.

3.3 Estimation Strategy

When estimating the effect of teaching practices on achievement, endogeneity bias may arise for different reasons. The estimated coefficients could be confounded by biases due to systematic self-selection and sorting of students and teachers to each other and/or to specific schools. For instance, if students with particularly positive unobserved characteristics such as high ability systematically select into schools with a large share of teachers who employ engaging teaching practices, any 'naïve' OLS estimate of the effect of these teaching practices on achievement would be upward biased.

One way of circumventing the endogeneity problem stemming from unobserved individual factors such as ability, lagged achievement, family background, and motivation is to estimate within-student between-subject models. This has been done by a number of economists (Bietenbeck, 2014; Clotfelter et al., 2010; Dee, 2005; Lavy, 2015; Schwerdt and Wuppermann, 2011). This procedure rules out bias due to unobserved individual factors,

[46] There is a rather large literature on the effects of computer use in classrooms, which, however, is not directly relevant for this article. Recent contributions from this strand of literature suggest that ICT use in classrooms is not per se good or bad, but depends on how computers are used and for what tasks (Falck et al., 2015; Lorena Comi et al., 2017).

because all the variation in these models stems from performance differences of the same individual in different subjects and their (systematic) association with differential input factors in these subjects. Based on this approach, I examine whether differences in achievement are systematically related to differences in teachers' use of engaging teaching practices in math, science, and reading. This is made possible by the fact that many students face different teachers in some or all of the three subjects. The fact that I am using information on three different subjects provides me with extra variation as compared to using just two subjects (mostly math and science) as is usually done in empirical research on education.[47] The basic idea for identification is that student, teacher, and school characteristics are constant across subjects except for differences in the frequency that teachers use engaging teaching practices and differences in all control variables.

Based on this identification strategy, I estimate education production functions of the following form:

$$A_{ijsk} = \alpha_i + \beta ESL_{jsk} + \gamma X_{is} + \delta T_{jsk} + \eta S_s + \tau_j + \xi_s + \varepsilon_{ijsk}, \tag{3.1}$$

where A_{ijsk} is the achievement of student i with teacher j in school s and subject k, ESL_{jsk} is teacher j's score on the ESL scale in school s and subject k, X_{is} is a vector of control variables pertaining to the personal background of student i in school s, T_{jsk} is a vector of covariates related to the personal background and teaching characteristics of teacher j in school s and subject k, and S_s is a vector of characteristics of school s.[48] The coefficient β is the main parameter of interest. τ_j and ξ_s represent unobserved characteristics of the teachers and the schools, while ε_{ijsk} is an idiosyncratic error term. Importantly, α_i is a student fixed effect that drops out of the within-student models. This fixed effect captures the effects of a

[47] Lavy (2015) also uses three different subjects. He examines the effect of instruction time on achievement among 15-year-olds.

[48] Note that I rule out within-school variation in class size, as classroom composition usually does not differ across subjects in primary school in Germany. This claim can be backed up by the data: Pairwise correlations between teacher-reported class sizes show coefficients of more than .98 in all three cases. Therefore, it seems safe to assume that the remaining variation is due to measurement error and, more generally, not relevant for the sake of this estimation. Practically, I am using teacher-reported class size in science as a proxy for all class sizes, as there are the fewest missing values in this variable.

student's family background, his or her prior educational career, innate ability, motivation, and other constant personality-related factors. Due to the student fixed effect, all general individual background factors that are observed in the data, denoted by X_{is}, also leave the model. Note that by controlling for a student fixed effect, I also control for school-level factors, as every student is only observed in one school. For that reason, the terms ηS_s and ξ_s drop out of the equation, too. Thus, the within-student models allow me to control for a wide range of student and school characteristics and their interactions that may cause bias in the estimations. Such bias could arise if there is a correlation between (unobserved) general school quality and the use of engaging teaching practices by teachers employed at this school. If teachers who spend a lot of time trying to engage their students systematically select into 'good' schools, upward bias will be introduced. The bias would be even stronger if high ability students were to self-select into these schools, too. However, a negative bias is also conceivable, for instance if some teachers who frequently use engaging teaching practices are at the same time keen on helping disadvantaged children and, as a result, sort into more "problematic" schools. Finally, it is imaginable that teachers adjust their behavior according to the group of students they are facing. For example, teachers may more frequently resort to encouraging their students and getting them involved in the subject matter when dealing with a group of less motivated students. This would also be a cause of downward bias in the OLS estimates. It is a priori unclear what kind of bias (upward or downward) should be expected. In any event, the student fixed effect effectively ensures that none of the above is a problem in the present study.

However, there are some issues in connection with my identification strategy that warrant mention. First of all, the effect captured by the coefficient β is "net" of any spillovers from one subject to another (i.e. if a student 'imports' his or her higher engagement triggered by teacher actions in subject A to subject B).

Second, a threat to my identification strategy could be student sorting to schools and teachers by subject-specific ability. Positive bias could result if students with high ability in math systematically chose schools in which the math teachers apply more engaging teaching practices. For this to happen, however, there would have to be clear differences in subject-specific ability between students. It is unclear to what extent this is true. For example, Clotfelter et al. (2010) provide evidence that academic ability is highly correlated across subjects. Even if significant subject-specific ability differences existed, a number of additional preconditions would have to be fulfilled so that my identification strategy would be threatened. First, parents would have to have prior knowledge about the specific strengths and weaknesses of their offspring. Second, teaching practices would have to systematically differ between subjects within the same school. And third, parents would have to have information about how teaching practices differ within schools. While the first condition may hold, it seems unlikely that all three conditions are met for a significant share of students. This notwithstanding, I can partially take care of the problem with the available data. Practically, I use a control variable in the empirical estimations that indicates whether or not a school suffers from a shortage of teaching materials in each of the three subjects. The idea behind this approach is that systematic differences in teaching practices within schools most likely occur in schools that specialize in certain subjects. Such schools, in turn, should be less likely to suffer from shortages of teaching materials in that subject.

A third concern would be systematic within-school sorting of students to teachers. This is less of a problem for primary school students than for secondary school students, however, as there are very few electives in primary schools. Furthermore, for such sorting to happen, the criteria outlined in the previous paragraph would have to be met, too, i.e. knowledge about subject-specific ability, subject-differences in teaching practices within schools, and information about the latter. One instance in which such information could exist is after children have started school, i.e. after first, second, or third grade. If they then switch to a different classroom, sorting could theoretically take place. I can deal with this problem by stratifying my sample according to good proxies of whether or not sorting is likely in a school. For instance, I know the total number of students in grade 4 in every school. By splitting the sample into smaller schools, which in many cases have only one class per grade, and schools with more classes, I can see whether any effects are concentrated among the larger schools that offer more room for within-school tracking. I also observe how much emphasis is given by schools to academic success. I assume that sorting into special ability classes is more likely in these schools than in others. Overall, the results I obtain from these stratifications are very similar to the baseline results. This suggests that within-school sorting by ability is not a likely cause of bias.

Fourth, my approach assumes that β is the same in all three subjects. This is in line with the theory laid out at the beginning of this chapter that student engagement should be a subject-independent input in educational production. However, in the latter part of the chapter, I will be able to relax this assumption and obtain subject-specific coefficients by estimating correlated random effects models.

Fifth, while it is true that my estimates are stripped of any unobserved individual and school-level heterogeneity, they could still be contaminated by non-random sorting of teachers into teaching practices. This challenge is faced by virtually all studies that deal with teaching practices and are not based on randomized controlled trials. Any bias introduced due to teacher sorting would be captured by the term τ_j in Equation (3.1). In practice, such sorting could arise if teachers with more favorable unobserved characteristics such as motivation or pedagogical skills use more engaging teaching practices. In that case, any positive effect of such teaching practices would be over-estimated due to unobserved teacher traits. In order to minimize this risk, I include a large set of teacher characteristics and teacher behavior variables as controls. In this respect, the expansion of teacher- and teaching-related information that has come with the 2011 wave of the TIMSS and PIRLS studies is of great value to me. It is further reassuring that Kane et al. (2011) find empirical evidence in teacher fixed effects estimations (with fewer teacher controls) that unobserved sorting of teachers into teaching practices is likely not a big issue in a similar setup. However, a closer look at the data is certainly warranted. Starting from the well-established idea that the amount of selection on observables provides some guidance to the magnitude of selection on unobservables, in Table B3.1 in Appendix B I provide estimates of the correlations between observable teacher characteristics and teachers' scores on the ESL scale. 10 out of 21 teacher controls turn out to be significantly related to the intention to engage students. So there does seem to be some systematic relation between teacher characteristics and certain teaching practices. To get an idea of how this affects my estimation outcomes, I will estimate different models with and without teacher- and teaching-related controls.[49] It is reassuring that the results do not change much depending on whether the set of covariates is included or not.

[49] Falck et al. (2015) follow the same procedure in an analysis of the effects of computer use in classrooms on achievement and find that the results do not change significantly.

3.4 Data

3.4.1 The TIMSS and PIRLS Studies

TIMSS and PIRLS are large-scale international assessment studies dealing with the educational achievement of fourth- and eighth-graders. TIMSS tests the knowledge and skills of students in math and science, while PIRLS is dedicated to the subject of reading. TIMSS is the ‘older’ study among the two. The first wave of testing was conducted in 1995. Subsequently, the study has been carried out every four years, i.e. in 1999, 2003, 2007, 2011, 2015, and 2019. PIRLS was first established in 2001 and has since been conducted every five years, i.e. in 2006, 2011, and 2016. Thus, so far the year 2011 has been the only occasion on which the two studies have coincided. As a result of this special timing, several countries decided to sample TIMSS and PIRLS together – at least among fourth-graders. The resulting dataset comprises information on 34 different countries and 3 benchmarking entities and allows researchers to analyze achievement of primary school students in three different subjects. In this study, I focus on country information on Germany. Here, a total of 4,067 students that are representative of the population of fourth-graders in the country were sampled (Bos et al., 2012a; 2012b).

Both TIMSS and PIRLS are administered by the International Association for the Evaluation of Educational Achievement (IEA), an organization first established in 1958 with vast experience in monitoring educational processes and outcomes. TIMSS and PIRLS apply a two-stage stratified sampling design. In the first stage, participating schools are chosen, and in the second stage, classes within these schools are selected. Stratification in TIMSS and PIRLS takes into account regional differences, school-type differences, level of urbanization, socio-economic indicators, and school performance on national examinations (Joncas and Foy, 2013). Testing in 2011 was carried out on two consecutive days; in half of the schools, students started with the TIMSS questionnaire on the first day and in the other half, students answered the PIRLS questionnaire first. The TIMSS assessment framework was organized around two different dimensions in 2011: content and cognition. The content section focused rather closely on what students should have learned in their curricula. In mathematics, this section contained questions related to numbers, geometric shapes, and measures, while in science, it comprised life science, physical science, and earth science.

The cognitive section put more emphasis on applying knowledge and reasoning. Generally, questions were split about evenly into multiple-choice and open-response. In total, the TIMSS questionnaire encompassed 175 items in math and 217 in science. The PIRLS assessment framework also focused on two different sections: reading for literary purposes and reading to acquire and use information. Within each of these sections, four comprehension processes were assessed: retrieving, inferencing, integrating, and evaluating. The text passages encompassed around 800 words with 13 to 16 questions underneath. PIRLS 2011 comprised a total of ten passages (five for each section), resulting in 135 questions (Martin and Mullis, 2013). To obtain as much information about the students' learning environment as possible, in addition to the actual tests, background questionnaires were administered to students, their parents, teachers, and school principals (Bos et al., 2012a, 2012b).

For my main variable of interest that measures the use of engaging teaching practices I exploit information from the teacher questionnaires. In 2011, TIMSS and PIRLS divided their teacher questionnaires into general questions that were answered by all teachers independent of the subject as well as subject-specific questions. The former are most useful for the purpose of my three-subject comparison. Among other components, the so-called Engaging Students in Learning Scale was introduced in this section of the questionnaires (Mullis, Martin, Foy, and Arora, 2012). The scale is inspired by work done by McLaughlin et al. (2005), who introduced the concept of student content engagement (Martin et al., 2012; Mullis et al., 2012a; 2012b). The ESL scale is based on teacher self-reports on specific classroom actions and is based on a six-item instrument. Specifically, teachers were asked how often they (1) summarize what students should have learned from the lesson, (2) relate the lesson to students' daily lives, (3) use questioning to elicit reasons and explanations, (4) encourage all students to improve their performance, (5) praise students for good effort, and (6) bring interesting materials to class. All questions could be answered on a four-point scale ranging from 'every or almost every lesson' over 'about half the lessons' and 'some lessons' to 'never or almost never' (IEA, 2011a, 2011b). Using item response theory, the raw data were transformed into the ESL scale by the IEA. The scale is constructed such that the mean of all participating countries is 10 and the standard deviation 2 (for more detailed information, see Martin et al., 2012). Note that scores on the ESL scale are constant within teachers, as teachers do not make statements on the use of engaging teaching practices by subject. Note also that the ESL scale is distinct from measures of modern or traditional teaching methods in that almost all items could be used in connection with both group-based modern approaches and lecture-style traditional teaching. Finally, it is worth mentioning that data based on teacher self-reports have previously been used in several studies (see e.g. Aslam and Kingdon, 2011; Hidalgo-Cabrillana and López-Mayan, 2015; Schwerdt and Wuppermann, 2011).

3.4.2 Sample Selection and Descriptive Statistics

My full sample comprises 4,067 students in 205 classes and 197 schools. However, in order to estimate the desired effect, I have to apply certain restrictions. First, I have to limit the analysis to students who have no more than one teacher per subject. That way, every student can be uniquely linked to exactly one teacher in math, one teacher in science and one teacher in reading. I thereby lose 135 students. Second, I consider only those students whose teachers have valid information on the ESL scale, which means that a further 106 students are excluded. And finally, I only keep students who participated in the achievement tests in all three subjects, which eliminates 413 students.[50] My final estimation sample consists of 3,413 students in 171 classes in 170 schools. This translates into 10,239 student-subject observations. Out of the 3,413 individual students, 1,684 students are taught by the same teacher in all three subjects, 1,024 students have the same teacher in science and reading but not in math, 434 students have the same teacher in math and reading but not in science, 190 students have the same teacher in math and science but not in reading, and 81 have different teachers in all three subjects. I leave the 1,684 students who have the same teacher in all subjects in the sample because of the valuable information on control variables that they provide. The large number of students with the same teacher is partly a result of the fact that in many cases math and science are taught as a single subject. In the robustness section, I demonstrate that including or excluding these students does not alter my results.

[50] In order not to lose too large a number of observations, I impute missing values on control variables by setting them to the respective mean and adding a dummy variable taking on the value 1 if the value was generated that way. In the case of dichotomous controls, I simply add a category for missing and use two dummies in the estimations with missing as the reference.

Table 3.1: Summary statistics of ESL scale by subject

	Mean	Std. dev.	Min.	Max.	*N*
Overall	8.74	1.56	2.95	13.27	10,239
Math	8.73	1.61	2.95	13.17	3,413
Science	8.74	1.57	4.57	13.27	3,413
Reading	8.75	1.51	4.57	13.27	3,413

Source: TIMSS/PIRLS 2011. Author's estimations.

Table 3.1 provides summary statistics of the ESL scale by subject. The teacher values on the scale range from 2.95 to 13.27, have a mean of 8.74 and a standard deviation of 1.56. This indicates that teachers in Germany make on average less use of engaging teaching practices than teachers in other countries. Generally, very few teachers state that they use the techniques in question never or almost never. Table B3.2 in Appendix B shows the means of the TIMSS and PIRLS achievement scores and the ESL scale as well as standard deviations both within students and between students. The mean test score for students in Germany is 533.8, well above the international centerpoint of 500 that was set as the mean achievement value in the first TIMSS and PIRLS studies. The standard deviation of test scores between students is 64.7, while the within-student standard deviation is about half as large at 31.5. That means there is substantial variation within students that can be explained in the regressions. The ESL scale has a between-student standard deviation of 1.36 and a within-student standard deviation of 0.56. To make the data more comparable across subjects and facilitate the interpretation of the results, I standardize both the achievement scores in math, science and reading to have mean 0 and standard deviation 1 and the teacher scores on the ESL scale.[51] For information on the remainder of the variables used in the empirical estimations refer to Table B3.3 in Appendix B.

[51] In the analysis, I use the first plausible value for all subjects. Each participant in TIMSS and PIRLS gets a total of five plausible values describing his or her performance. Plausible values are used to correct for different degrees of difficulty in the exercises, as not all students answer the exact same questions.

3.5 Results

3.5.1 Main Results

Table 3.2 reports the estimated coefficients of the effect of engaging teaching practices on individual achievement. All regressions contain subject fixed effects and are weighted by probability weights as supplied in the TIMSS dataset. Columns 1 and 2 present results of pooled OLS models, while columns 3 and 4 report estimates based on student fixed effects specifications. The coefficients in the OLS models are negative and borderline significant. While the estimate reported in column 1 from a model containing only personal and school background control variables reaches statistical significance at the 90 percent level, the coefficient from the full model that includes comprehensive information on teachers, classrooms, and teaching practices narrowly fails to reach significance (see column 2). These estimates, which suggest that engaging teaching practices may have a negative effect on achievement, are potentially biased by all sorts of student and teacher self-selection into schools and classrooms. In fact, when considering the student fixed effect models and especially my preferred specification in column 4, statistical significance disappears and the point estimate is equal to zero. In these models, student self-selection should play no role. Against the background of potential sorting of teachers into different teaching practices, it is reassuring that the inclusion of teacher- and teaching-related control variables in column 4 compared to column 3 does not significantly alter the results. If at all, they make the results more positive, which suggests that the causal effect of engaging teaching practices may be larger (i.e. positive) but probably not smaller (i.e. negative) than displayed in Table 3.2. This implies that I can probably rule out any harmful effects of the use of engaging teaching practices on achievement.

The basic conclusion to be drawn from the main results is that the frequency of the use of engaging teaching practices does not affect achievement. The difference in the results between the OLS models and the student fixed effects models suggests that there is some negative sorting of either students to teachers or teachers to students and/or schools in Germany. For instance, it may be that parents of less motivated or low-ability children intentionally send their offspring to schools that are known for their engaging teaching practices. The

Table 3.2: Estimated effect of engaging teaching practices on student achievement

	OLS		Student FE	
	(1)	(2)	(3)	(4)
(A)				
ESL	-.041*	-.030	-.005	-.001
	(.022)	(.021)	(.012)	(.012)
(B)				
ESL	-.040	.037	-.001	.039
	(.126)	(.139)	(.060)	(.088)
ESL-squared	-.000	-.002	-.000	-.001
	(.004)	(.005)	(.002)	(.003)
Subject FE	yes	yes	yes	yes
Personal and school characteristics	yes	yes		
Teacher and teaching characteristics		yes		yes
N	10,239	10,239	10,239	10,239

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. This table shows regressions of students' z-standardized achievement scores on teachers' z-standardized values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. In the upper panel, only the level of teachers' values on the ESL scale are considered. In the lower panel, the levels and squared terms of values on the ESL scales enter the models. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Source: TIMSS/PIRLS 2011.

negative coefficients in the OLS models could also be the result of reverse causality, i.e. if teachers of worse students more often resort to engaging teaching practices than other teachers. This makes intuitive sense, as especially low-performing students may need and receive additional teacher support. For high-performing students, it is imaginable that teachers substitute engaging teaching practices for other classroom actions, since students follow the course content in any event. While there is no natural counterpart to engaging teaching practices, the most likely alternative would be giving additional exercises, as instructional strategies to raise engagement mostly focus on repetition, summarizing, questioning, encouraging and praising, which, to some extent, crowds out additional content. Thus, engaging students may be related to more intense study of certain material at the cost of additional material.

Since it is likely that a mix of different classroom actions produces the highest achievement, I report estimates for models that allow for non-linearities in the lower panel of Table 3.2. Practically, I add a squared term of the standardized teacher score on the ESL scale to the models. However, no estimate from these models turns out significant. This may have something to do with the way the questions are framed in the questionnaires. Teachers are asked whether they apply the engaging teaching practices in ‘every or almost every lesson’, ‘about half the lessons’, ‘some lessons’, or ‘never’. This is different from asking whether teachers try to engage their students *during* the whole lesson, half the lesson or less often, since it does not offer information on the actual time spent on engaging students. If that were the case, a significant non-linear effect would intuitively be more likely to appear because higher values would necessarily mean more crowding out of alternative teaching practices such as giving additional exercises.

3.5.2 Heterogeneous Effects

A lack of significant effects in the full sample does not preclude the possibility that certain subgroups of students may still be affected positively by engaging teaching practices. For instance, in an article on modern and traditional teaching practices, Lavy (2015) finds that children from different socio-economic backgrounds are quite differently affected by teachers’ classroom actions. In other words, a specific teaching practice may be good for some students but bad for others, simply because different groups of students have different needs. Heterogeneous effects could also be related to differences in the ‘baseline’ engagement of different groups of students. If there are many students in a particular group who are engaged during the whole lesson in any case, further time spent on engaging them should not be advantageous. In all likelihood they would even be negatively affected because their teachers’ attempts to further engage them crowd out alternative actions such as giving additional exercises. While this is an extreme example, differences in the average baseline engagement could be expected between students from different socio-economic backgrounds and between boys and girls. For instance, already in primary school, boys are often found to be less engaged in schooling matters than girls (McCoy et al., 2012). Socio-economic status could matter if students from higher socio-economic backgrounds have learned a more "pro-education" attitude from their parents. However, empirical evidence on this is not conclusive (for an overview see Shernoff, 2013). Finally, one could expect differences between children who speak German with their parents and those who do not, as the latter may need to be addressed differently in class.

Table 3.3: Estimated effect of engaging teaching practices on standardized test scores for different subgroups, student fixed effects models

	Socio-economic background		Gender		Language mostly spoken at home	
	High (1)	Low (2)	Boys (3)	Girls (4)	German (5)	not German (6)
(A)						
ESL	-.015	.046**	.003	.006	.006	-.014
	(.015)	(.021)	(.016)	(.014)	(.014)	(.026)
(B)						
ESL	-.093	.383**	.135	-.037	.130	.056
	(.075)	(.156)	(.129)	(.086)	(.085)	(.247)
ESL-squared	.003	-.012**	-.005	.002	-.004	-.002
	(.003)	(.005)	(.005)	(.003)	(.003)	(.009)
Subject FE	yes	yes	yes	yes	yes	yes
Teacher and teaching characteristics	yes	yes	yes	yes	yes	yes
N	4,314	3,297	5,124	5,115	7,401	1,866

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. This table shows regressions of students' z-standardized achievement scores on teachers' z-standardized values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. In the upper panel, only the level of teachers' values on the ESL scale are considered. In the lower panel, the levels and squared terms of values on the ESL scales enter the models. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Source: TIMSS/PIRLS 2011.

Table 3.3 presents results of subgroup-specific estimations. Again, the upper panel deals with linear analyses while the estimates in the lower panel come from specifications that allow for non-linearities. As can be seen in column 2, I find a positive effect for children from low socio-economic backgrounds.[52] The magnitude of the effect is small to modest: A one-standard-deviation-increase on the ESL scale raises test scores in math, science, and reading by 4.6 percent of a standard deviation of the test score distribution. This is equivalent to approximately three points on the achievement tests. This finding suggests that engaging students in learning can yield societal gains in terms of greater equality of opportunity. This is important in Germany, where intergenerational educational mobility is generally found to be low (Heineck and Riphahn, 2009). Still, the lower panel of column 2 suggests that the observed effect may not be linear along the distribution of values on the

[52] Children are defined as having a low socio-economic background if neither of their parents has a post-secondary degree. All others are classified as having a high socio-economic background.

ESL scale. While the level coefficient is positive, significant and very large, the coefficient estimate of the squared term is significantly negative. However, the calculated turning point of 15.96 is well outside the data range (recall that the data are standardized). Thus, while there may be decreasing returns to the use of engaging teaching practices, there does not seem to be a relevant non-monotonic relationship in practical terms. Apart from children from low socio-economic backgrounds, no other subgroup seems to benefit from engaging teaching practices. Thus, for children from high socio-economic backgrounds, boys, girls, children who mostly speak German at home and those who do not, the null results of the full sample are confirmed.

Generally, it is important to note that relative subgroup differences are less likely to be a consequence of sorting on unobserved teacher characteristics than the overall results. This is the case because even if there is systematic sorting of teachers into certain teaching practices, such sorting will affect all subgroups in the same way. The only concern in this case would be subgroup-specific sorting. For such subgroup-specific sorting to happen, there would, for instance, have to be unobserved teacher characteristics, which are especially beneficial or detrimental to just a particular subgroup of students and teachers would have to sort into certain teaching practices along these characteristics – something that has been deemed unlikely by other researchers in the past (see e.g. Lavy, 2015).

3.5.3 Robustness of the Results

In the following section, I present analyses that (1) underscore the robustness of my results to alternative specifications and definitions of treatment and (2) support their causal interpretation. As a first robustness check, I exclude all classes with less than 16 students from the analysis.[53] This reduces my sample by 324 observations to 9,915 student-subject pairings. The reason for excluding those very small classes is that the general classrooms dynamics may be very different in them as opposed to larger classes. Most importantly, in larger classes the potential for disturbances and interruptions increases, which may render

[53] I chose 16 students as the cut-off point after visual inspection of the frequency distribution of class sizes. It becomes much denser starting with classes of 16 students. In total, there are 246 students in classes with 16 students, while there are only 90 students in classes with 15 students.

Table 3.4: Robustness of the estimated effect of engaging teaching practices on standardized test scores for different subgroups, student fixed effects models

	Full Sample	Socio-economic background		Gender		Language mostly spoken at home	
		High	Low	Boys	Girls	German	not German
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(A)							
Excluding small classes (< 16 students)	-.001 (.012)	-.017 (.015)	.046** (.022)	.005 (.016)	.003 (.014)	.007 (.014)	-.013 (.026)
N	9,915	4,203	3,207	4,953	4,962	7,164	1,794
(B)							
Only students with variation in teachers	.003 (.012)	-.007 (.015)	.054** (.023)	.006 (.017)	.010 (.014)	.009 (.014)	-.016 (.025)
N	5,187	2,253	1,548	2,526	2,661	3,918	870
(C)							
Treatment based on factor analysis	-.005 (.016)	-.026 (.017)	.063** (.025)	-.003 (.019)	.004 (.019)	.012 (.018)	-.033 (.035)
N	10,122	4,250	3,254	5,072	5,050	7,313	1,850

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. This table shows regressions of students' z-standardized achievement scores on teachers' z-standardized values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Source: TIMSS/PIRLS 2011.

the task of engaging students more important. However, one could also imagine that it is easier for teachers to 'reach' their students with their attempts to engage them in small groups. In any case, the estimates presented in panel A of Table 3.4 suggest that different classroom dynamics in small classes do not drive my results. The finding of no overall effects and modest gains for students from low socio- economic backgrounds holds. In fact, the point estimate of the coefficient for students from low socio-economic backgrounds is exactly the same as in the full (sub)sample regression (0.046).

As a second robustness check, I restrict my sample to those students that are not taught by the same teacher in all three subjects. This cuts my sample roughly in half to 5,187 students. However, it does not have a significant effect on the results, which are reported in panel B of Table 3.4. As compared to the estimates in Tables 2 and 3, they seem only marginally more positive. For example, the estimated benefit from a one-standard-deviation-increase in the use of engaging teaching practices for students from low socio-economic backgrounds is now 5.4 percent of a standard deviation of the test score distribution and thus 17 percent higher than in column 2 of Table 3.3. Apart from this significant effect, the general pattern of no significant effects in the full sample and for all other subgroups holds.

A third robustness check aims at manipulating my main independent variable, the score on the ESL scale. Recall that the IEA provides a ready-made scale based on item response theory in the dataset. I use the information from the six items to construct an alternative ESL scale based on factor analysis. First of all, it is encouraging that all items load on one factor. Visual inspection of the screeplot confirms this result (see Figure B3.1 in Appendix B). The results of the regressions with the resulting factor score as principal regressor are provided in panel C of Table 3.4.[54] Once more, the pattern of all previous regressions is confirmed: Only children from low socio-economic backgrounds stand to gain from engaging teaching practices. The point estimate suggests that a one-standard-deviation-increase in the ESL scale is associated with an increase in test scores of 6.3 percent of a standard deviation. This effect is 37 percent larger than in the baseline model and 17 percent larger than in the specification without students who are taught by the same teacher in all three subjects.

The fourth robustness check concerns the question of whether my results are contaminated by subject-specific sorting and self-selection in some schools. It makes use of information from the school background questionnaire of the TIMSS and PIRLS studies. School principals are asked about the total number of students enrolled in grade 4 in their school. It turns out that the smallest school has only 6 students in grade 4, the largest 158. I use this information to separately estimate models for large and small schools. The cut-off point for being a small school is 31 students, which is equivalent to the largest number of students in one classroom in my data. It is also equivalent to the largest maximum class size rule in place in German primary schools at the time (in the state of Baden Wurttemberg). This gives me one group of 1,311 students in small schools and a significantly bigger group of

[54] Note that I multiplied the factor score by -1 due to the way the questions are coded in the teacher questionnaire where a higher value signifies *less* frequent use of the technique in question. This manipulation makes the results easier to understand and compare to the other estimates.

8,592 students in large schools. I expect that this procedure lends further credibility to my identification strategy, as there should be much less room for tracking in small schools. In fact, most schools should only have one classroom per grade as almost all German states operated maximum class size rules at the time that were set at total grade enrollment multiples of between 28 and 31 (Kultusministerkonferenz, 2007).[55] In all cases where total enrollment did not pass this threshold, there would be only one class per grade 4. If there were significant within-school sorting on unobservables, I would expect that the results of the two groups starkly differ from one another. More precisely, if there were positive sorting of students to teachers by subject-specific ability and teaching practices, one would expect the estimate for large schools to be larger than the one for small schools (and vice versa if there were negative sorting). As the sample size gets rather small as a result of splitting the sample, I am not able to perform subsample analysis, e.g. for children from different socio-economic backgrounds. However, my main interest here lies in finding out whether or not my analysis *generally* suffers from omitted variable bias. The lessons from this exercise are as follows: First, the results shown in columns 1 and 2 of Table 3.5 are encouraging in the sense that neither the estimate for small schools nor the estimate for large schools turns out significant. This suggests that my results are not severely biased by sorting within schools. Second, if one were to disregard significance and only look at effect sizes, one would have to conclude that the effect is more positive in small schools. Clearly, to the (limited) extent that there is sorting within schools, it seems to be negative. This pattern confirms my previous results and strengthens the conclusion that a negative association between engaging teaching practices and achievement can be ruled out while a small positive effect cannot entirely be precluded. The null result depicted in column 4 of Table 3.2 could therefore be interpreted as a lower bound on the actual effect. Importantly, the results of this robustness test are not sensitive to picking any cut-off value for small schools between 28 and 31.

[55] This refers to maximum class size rules for the school year 2007/2008, which is when most of the students in the sample entered primary school.

Table 3.5: Estimated effect of engaging teaching practices on standardized test scores by size of grade 4 and emphasis on academic success, student fixed effects models

	Grade size		Emphasis on academic success	
	Large (1)	Small (2)	(Very) High (3)	Medium (4)
(A)				
ESL	-.009	.043	.012	.059
	(.014)	(.030)	(.018)	(.074)
N	8,592	1,311	6,825	2,952

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. This table shows regressions of students' z-standardized achievement scores on teachers' z-standardized values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Source: TIMSS/PIRLS 2011.

To further underscore this claim, I performed a fifth robustness check, which again splits the sample by schools that are more likely to sort students and schools that are less likely to do so. Again, the data is provided by school principals, who are asked a total of five questions about mainly teacher, student, and parent expectations regarding academic success in their schools.[56] The answers are used by the IEA to construct a so-called School Emphasis on Academic Success Index. This index has three different categories: very high emphasis, high emphasis, and medium emphasis. Note that more than 99 percent of all students visit a school that falls into one of the two latter categories. For practical reasons, I added all students going to a school with a very high emphasis on academic success to those visiting a school with a high emphasis. The resulting dichotomous indicator shows that more than two thirds of all students for whom there is information go to a (very) high emphasis on academic success school. My prior is that the former should be more prone to forming special ability groups and classrooms and, therefore, to endogenous sorting and selection. If the results were driven by such sorting, they should be different from the rest among these schools. Yet, both estimates turn out insignificant again (see columns 3 and 4 in Table 3.5). Once more, however, the point estimate is larger in the sub-sample where there should be less sorting (medium emphasis on academic success), which gives further credibility to the interpretation of my main result as a lower bound on the actual effect.

[56] The five items belong to question 12 of the school context questionnaire, which had to be answered on a five-point-scale that ranges from very high to very low.

3.5.4 Subject Differences

So far, I have assumed that trying to engage students in learning is equally effective (or ineffective) in math, science, and reading. In reality, this need not be the case. For this reason, I estimate models based on the three possible samples that include only two of the three subjects. The results of this exercise are presented in Table B3.4. Again, the estimated effect of engaging teaching practices on achievement does not turn out significant in any of the models.

The two-subject models do not provide definitive proof for the hypothesis that engaging students has the same effect in all subjects. A more refined method to do this is to estimate correlated random effects models. By explicitly modeling the potential correlation of unobserved heterogeneity stemming from individual- and school-level factors, such models allow me to relax the assumption of constant coefficients across subjects and to estimate subject-specific specifications. For details on this strategy, consider Appendix A. The results indicate that subject differences can, in fact, not entirely be ruled out. Column 1 in Table 3.6 shows a weakly significant positive effect of engaging teaching practices on achievement in reading. This effect is slightly smaller than the estimated positive effect for children from low socio-economic backgrounds from the fixed effect models. Specifically, a one-standard-deviation-increase in the ESL scale is associated with an increase in test scores of 4.3 percent of a standard deviation. In the other two subjects, the estimates are close to zero and not significant, which confirms the results of the fixed effects models.

The results for different subgroups are broadly in line with the results of the fixed effects models. The coefficients for children from high socio-economic backgrounds are negative in all models and only in the case of math weakly significant. The null hypothesis of equal coefficients in all subjects cannot be rejected. Children from low socio-economic backgrounds are again much more positively affected. However, this is only true for the two subjects of math and reading where the estimated effects are one-and-a-half (mathematics) to two times (reading) the size of the estimate from the fixed effects models. For boys and girls I find no significant effects in any of the models and the estimated coefficients are not significantly different from one another. This confirms my previous results. There are also

Table 3.6: Estimated effect of engaging teaching practices on standardized test scores by subject, correlated random effects models

	Full sample	Socio-economic background		Gender		Language mostly spoken at home	
		High	Low	Boys	Girls	German	not German
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math	.010 [0.22]	-.054* [2.82]	.072** [4.76]	.001 [0.00]	.010 [0.12]	.029 [1.68]	-.071 [0.67]
Science	-.004 [0.04]	-.027 [0.94]	-.006 [0.03]	-.000 [0.00]	-.008 [0.12]	.013 [0.47]	-.111 [1.51]
Reading	.043* [2.98]	-.047 [1.62]	.100** [5.94]	.049 [2.62]	.032 [1.08]	.055* [3.58]	.048 [0.27]
N	3,413	1,438	1,099	1,708	1,705	2,467	622
$\beta^{math} = \beta^{science} = \beta^{read}$	[7.95]**	[2.42]	[8.61]***	[4.34]	[3.96]	[3.77]	[11.72]***

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. χ^2 statistics in brackets. Correlated random effects models estimated by seemingly unrelated regressions (SUR). Regressions of students' z-standardized achievement scores on teachers' z-standardized values on the ESL scale. All regressions are weighted by the students' sampling probability. Standard errors are clustered at the classroom level. Control variables are listed in Table A.3. Source: TIMSS/PIRLS 2011.

no significant effects for children who do not speak German at home. The fact that the estimated coefficients differ from one another, nevertheless, is a result of the small sample size, which leads to rather imprecise estimates. For children who speak German at home I find a weakly positive effect in reading. However, it cannot be concluded that this coefficient statistically differs from the other coefficients.

All in all, the results of the correlated random effects models are qualitatively similar to the ones of the fixed effects models. Most importantly, they confirm the finding that mostly children from low socio-economic backgrounds are positively affected by engaging teaching practices. However, they also show that in some cases subject differences cannot entirely be ruled out. It emerges that in reading, the use of engaging teaching practices may be especially beneficial. While the theory outlined above would predict otherwise, this result can be reconciled with previous research by Tyler et al. (2010) and Kane et al. (2011). They find that engaging students in questions and discussions is particularly effective in reading.

A tentative explanation for this could be that reading results are to a lesser extent determined by general cognitive capabilities and more responsive to effort than science and math results. This interpretation is supported by a paper by Deary et al. (2007) who find that the association of cognitive ability with math achievement is much stronger than with English language achievement.

3.6 Conclusion

In this paper, I have investigated the effects of primary school teachers’ uses of engaging teaching practices on achievement as measured by standardized assessment studies. The object of my analysis was a nationally representative sample of fourth-graders in Germany. The use of engaging teaching practices was measured by the ESL scale as supplied by the IEA in connection with the TIMSS and PIRLS studies 2011. It is based on questions regarding how often teachers use questioning in class, bring interesting materials to the course, relate the course content to students' daily lives, give praise and encouragement and summarize the most important points of the lesson. The main finding is that engaging teaching practices yield modest achievement gains for students from low socio-economic backgrounds. In the full sample, no significant effects could be detected. However, I was able to virtually rule out any harmful effects of engaging teaching practices.

The identification strategy, which is based on a novel within-student between-subject approach in three different subjects, reliably rules out unobserved heterogeneity stemming from individual- and school-level characteristics. It has certain limitations regarding teacher sorting into specific teaching practices, which should be borne in mind when interpreting the results. Nonetheless, the relative position of students from low socio-economic backgrounds as compared to students from high socio-economic backgrounds is even more likely to reflect truly causal mechanisms than the full sample results. The reason for this is that any overall teacher-related bias would affect both groups in the same way unless sorting of teachers into teaching practices along unobserved characteristics is particularly beneficial or detrimental to certain subgroups.

From a policy perspective, the results of the present analysis can be understood as a possible vehicle to achieve greater equality of opportunity; especially since the gains for children from low socio-economic backgrounds are not offset by detrimental effects on children from high socio-economic backgrounds. From an efficiency perspective, one would also have to assess the costs of implementing more engaging teaching practices in schools across the country. However, especially for future teachers, these costs would probably not be prohibitively high, as they would mainly arise from slightly altering the focus of teacher training.

The results of this paper open up a number of fruitful avenues for future research. First of all, a lot remains unknown about what classroom actions are effective in conferring skills upon students. This is related to the question of teachers' time allocation between different teaching practices, as it is likely that a mix of different actions generates the best results. Secondly, this paper has shown that not all teaching practices need to be equally effective for all students. Against this background, more and deeper subgroup-specific analysis would be desirable. Of course, the feasibility of this hinges upon the provision of better data. For instance, the present work would have vastly benefited from subject-specific information on the use of engaging teaching practices, as this would have allowed within-teacher estimations, which in turn would have generated more definite conclusions on the causal nature of the results.

Appendix A: Correlated Random Effects Models

The educational production function given in Equation (3.1) can be slightly altered so that the effect of the use of engaging teaching practices on achievement is allowed to vary by subject. This yields:

$$A_{ijs}^k = \beta^k ESL_{js}^k + \gamma^k X_{is}^k + \delta^k T_{js}^k + \eta^k S_s^k + \varepsilon_{ijs}^k \qquad \text{(A3.1)}$$

$$\varepsilon_{ijs}^k = \alpha_i^k + \tau_j^k + \xi_s^k \qquad \text{(A3.2)}$$

The production function depicted in Equation (A3.1) can be estimated separately for each subject k. However, the drawback of subject-specific estimations is that the individual-level component α_i^k as well as the school-level component ξ_s^k of the total error term ε_{ijs}^k are additional sources of potential bias compared with student fixed effects models. One way of dealing with this problem is to estimate correlated random effects models that explicitly model the potential correlation of the unobserved heterogeneity with the observed inputs (see e.g. Ashenfelter and Zimmerman, 1997; Falck et al., 2015; Metzler and Wößmann, 2012; Piopiunik and Schlotter, 2012). This method was first proposed by Mundlak (1978) and later refined by Chamberlain (1982, 1984). In the spirit of Chamberlain's work, I model the potential correlation between the error components α_i^k and ξ_s^k and all observed inputs in a very general way:

$$\alpha_i^k = \zeta_1^k ESL_{js}^{mat} + \zeta_2^k ESL_{js}^{sci} + \zeta_3^k ESL_{js}^{rea} + \theta_1^k X_{is} + \rho_1^k T_{js}^{mat} + \rho_2^k T_{js}^{sci} + \rho_3^k T_{js}^{rea} + \iota_1^k S_s + \omega_i^k \qquad \text{(A3.3)}$$

$$\xi_s^k = \psi_1^k ESL_{js}^{mat} + \psi_2^k ESL_{js}^{sci} + \psi_3^k ESL_{js}^{rea} + \theta_2^k X_{is} + \rho_4^k T_{js}^{mat} + \rho_5^k T_{js}^{sci} + \rho_6^k T_{js}^{rea} + \iota_2^k S_s + \sigma_i^k \qquad \text{(A3.4)}$$

Here, the individual and school-level error terms depend on all subject-specific realizations of the independent variable of interest and all other covariates. In addition, the model contains the subject-specific error terms ω_i^k and σ_i^k. The intuition behind this way of modeling the error term is the empirical observation made in connection with panel data that if an error term ε_i is correlated with some explanatory variable X_{it} in period t_1, it will also be correlated with X_{it} in period t_2, where $t_1 \neq t_2$ (Chamberlain, 1984; Mundlak, 1978). Hence, the correlation between ε_i and X_{it} can be modeled for each t_{1-N} by using information on X_{it} from all other t_{1-N}. I make use of this observation and simply replace the time dimension *t* with the subject dimension *k*.

Plugging equations (A3.3) and (A3.4) into Equation (A3.1) for each k yields the following correlated random effects models, which are estimated by seemingly unrelated regressions via maximum likelihood:

$$A_{ijs}^{mat} = (\beta^{mat} + \zeta_1^{mat} + \psi_1^{mat})ESL_{js}^{mat} + (\zeta_2^{mat} + \psi_2^{mat})ESL_{js}^{sci} + (\zeta_3^{mat} + \psi_3^{mat})ESL_{js}^{rea} + e_{ijs}^{mat} \tag{A3.5}$$

$$A_{ijs}^{sci} = \left(\beta^{sci} + \zeta_2^{sci} + \psi_2^{sci}\right)ESL_{js}^{sci} + \left(\zeta_1^{sci} + \psi_1^{sci}\right)ESL_{js}^{mat} + \left(\zeta_3^{sci} + \psi_3^{sci}\right)ESL_{js}^{rea} + e_{ijs}^{sci} \tag{A3.6}$$

$$A_{ijs}^{rea} = (\beta^{rea} + \zeta_3^{rea} + \psi_3^{rea})ESL_{js}^{rea} + (\zeta_1^{rea} + \psi_1^{rea})ESL_{js}^{mat} + (\zeta_2^{rea} + \psi_2^{rea})ESL_{js}^{sci} + e_{ijs}^{rea} \tag{A3.7}$$

$$e_{ijs}^k = (\gamma^k + \theta_1^k + \theta_2^k)X_{is} + \left(\delta^k + \sum_{n=1}^{6} \rho_n^k\right)T_{js}^k + (\eta^k + \iota_1^k + \iota_2^k)S_s + \omega_i^k + \sigma_i^k + \tau_j^k \tag{A3.8}$$

The parameters of interest β^{mat}, β^{sci}, and β^{rea} cannot be directly observed from the regression results, as they are confounded by the coefficients ζ_n^k and ψ_n^k, which together reflect the size of the respective selection bias. The parameters β^k can be obtained by taking differences over $\beta^{mat} + \zeta_1^{mat} + \psi_1^{mat}$ and $\zeta_1 + \psi_1$for math, $\beta^{sci} + \zeta_2^{sci} + \psi_2^{sci}$and $\zeta_2 + \psi_2$ for science, and $\beta^{rea} + \zeta_3^{rea} + \psi_3^{rea}$ and $\zeta_3 + \psi_3$ for reading. This requires an implicit over-identifying restriction of correlated random effects models to hold, namely that $\zeta_n + \psi_n$ is the same across all subjects k, formally: $\zeta_n^{mat} + \psi_n^{mat} = \zeta_n^{sci} + \psi_n^{sci} = \zeta_n^{rea} + \psi_n^{rea} = \zeta_n + \psi_n$. Practically, I can impose this condition by restricting the sum of the coefficients ζ_n^k and ψ_n^k to be the same in the two respective "out-of-subject" equations. In the case of math achievement, this means restricting the term $\zeta_1^{sci} + \psi_1^{sci}$ in Equation (A3.6) (science achievement) to be equal to the term $\zeta_1^{rea} + \psi_1^{rea}$ in Equation (A3.7) (reading achievement). The resulting estimate is used to strip the bias $\zeta_1^{mat} + \psi_1^{mat}$ off the observable term $\beta^{mat} + \zeta_1^{mat} + \psi_1^{mat}$, thereby obtaining an estimate of β^{mat} that is unbiased by unobserved individual- and school-level heterogeneity. This procedure is carried out for all three subjects separately.

Appendix B: Additional Tables and Figures

Table B3.1: Pairwise correlations between observable teacher characteristics and teacher scores on the ESL scale

	Pearson's r
Objective characteristics	
Experience	0.24***
Sex	-0.00
Age	0.20***
Education	0.00
Field teacher	0.05
Interactions with other teachers	
Discuss how to teach a particular subject	0.20***
Collaborate in planning and preparing materials	0.05
Share teaching experiences	0.29***
Visit other classrooms to learn	0.12**
Work together to try out new ideas	0.17***
Job satisfaction	
Content with teaching profession	-0.07
Satisfied being a teacher at this school	-0.08
Had more enthusiasm when I began teaching	0.09
Do important work as a teacher	-0.06
Plan to continue as a teacher for as long as I can	0.07
Frustrated as a teacher	0.07
Relation to parents	
Individually discuss learning progress	-0.16***
Send home a progress report	0.01
Use of computers	
for preparation	-0.06
for administration	0.11*
for classroom instruction	-0.14**

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Source: TIMSS/PIRLS 2011.

Table B3.2: Descriptive statistics – test scores and ESL scale

	Test scores	ESL scale
Mean	533.8	8.73
SD between students	64.7	1.36
SD within students	31.5	0.56

Source: TIMSS/PIRLS 2011.

Table B3.3: Summary statistics of covariates

	Mean	SD		Mean	SD
Student and school controls			**Interactions with other teachers**		
Male	0.50	0.50	Discuss how to teach a particular subject**	2.37	0.87
Age (months)	124.3	6.05	Collaborate in planning and preparing materials**	2.41	0.79
Mostly German spoken at home*	0.73	0.44	Share teaching experiences **	2.50	0.90
Parent involvement in learning (once or twice a week vs. less)*	0.80	0.40	Visit other classrooms to learn**	1.13	0.40
School size (N students)	56.9	25.3	Work together to try out new ideas**	1.98	0.79
N Computers (in school)	14.9	9.22	**Job satisfaction**		
Avg. student background in school (3 categories)	2.02	0.64	Content with teaching profession***	1.58	0.63
N people in school area (6 categories)	3.93	1.58	Satisfied being a teacher at this school***	1.42	0.55
School emphasis on success (3 categories)	2.30	0.46	Had more enthusiasm when I began teaching***	2.58	1.04
School discipline and safety (3 categories)	1.62	0.56	Do important work as a teacher***	1.14	0.36
Teacher- and teaching-related characteristics			Plan to continue as a teacher for as long as I can***	1.59	0.81
Resource shortages (TIMSS/PIRLS scale)	10.62	1.56	Frustrated as a teacher***	3.38	0.69
Class size (N students)	21.7	3.84	**Relation to parents**		
Instructional time (minutes per week)	252.2	141.6	Individually discuss learning progress****	3.45	0.71
Experience (years)	18.9	12.5	Send home a progress report****	4.43	0.72
Female	0.87	0.34	**Use of computers**		
Age (1 = <30; 2 = 30-39; 3 = 40-49; 4 = >49)	2.94	1.05	for preparation (1=yes, 0=no)	0.97	0.16
Education (tertiary or not)*	0.89	0.31	for administration (1=yes, 0=no)	0.84	0.37
Field teacher (majored in subject)*	0.65	0.48	for classroom instruction (1=yes, 0=no)	0.76	0.43

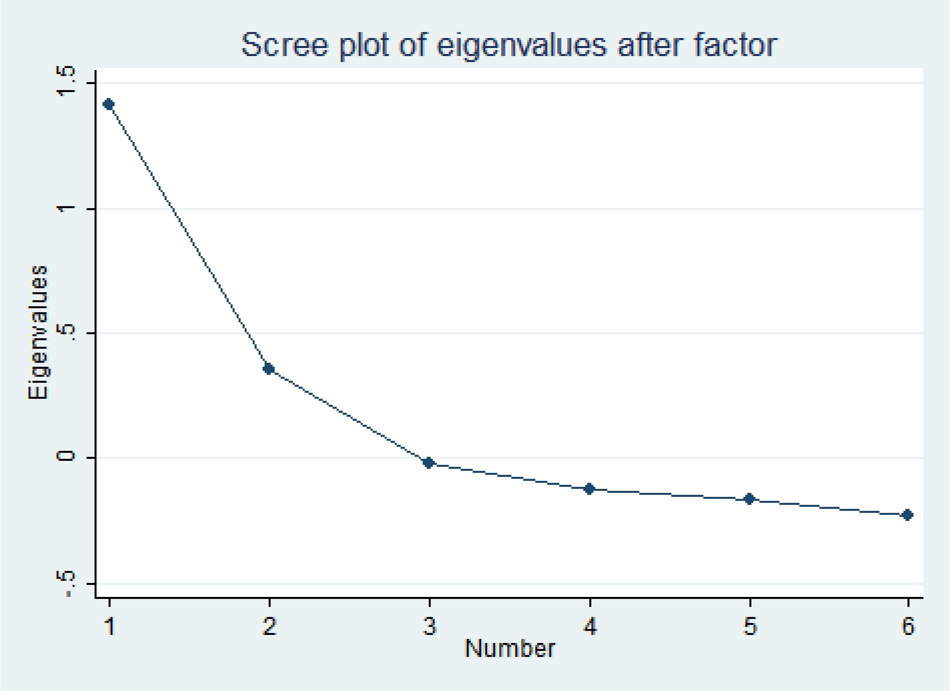
Notes: * Two dummies (yes/no) with missing as the reference. ** Variables are based on a 4-category scale with 1 equaling ‘agree a lot’ and 4 equaling ‘disagree a lot.’ *** Variables are based on a 4-category scale with 1 equaling ‘never or almost never’ and 4 equaling ‘daily or almost daily.’ **** Variables are based on a 5-category scale with 1 equaling ‘at least once a week’ and 5 equaling ‘never.’ Source: TIMSS/PIRLS 2011.

Table B3.4: Estimated effect of engaging teaching practices on student achievement; student fixed effects models, two subjects at a time

	Math + Science (1)	Math + Reading (2)	Science + Reading (3)
(A)			
ESL	.002	-.030	.034
	(.010)	(.025)	(.048)
Subject FE	yes	yes	yes
Teacher and teaching characteristics	yes	yes	yes
Number of observations	6,826	6,826	6,826

Notes: * p < 0.10, ** p < 0.05, *** p < 0.01. This table shows regressions of students' z-standardized achievement scores on teachers' z-standardized values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table B3.3. Source: TIMSS/PIRLS 2011.

Figure B3.1: Screeplot of eigenvalues after factor analysis on six items of engaging teaching practices



Chapter 4

Birth Cohort Size Variation and the Estimation of Class Size Effects

4.1 Introduction

Class size is one of the most important determinants of the costs of education as teachers' salaries by far comprise the largest share of public expenditures on education in most countries (OECD, 2019). At the same time, the empirical literature on class size effects is contentious and does not offer clear guidance as to what are the effects on student outcomes that class size changes entail. To identify these effects, a large part of the quasi-experimental literature exploits within-school variation in cohort size over time (see e.g. Hoxby, 2000; Leuven et al., 2008; Cho et al., 2012). These studies mostly find small or no class size effects, which contrasts with the available experimental evidence showing substantial class size effects (see e.g. Krueger, 1999; Krueger and Whitmore, 2001).[57]

This paper offers a potential explanation for this apparent puzzle. In school systems that allow students with insufficient academic skills to be held back a grade, we can show that class size estimates based on within-school variation in cohort size are upward biased because of a mechanical relationship between the initial size of a cohort and the student composition in higher grades. This bias has been ignored to date and helps to explain why studies using within-school variation in cohort size generally find less negative class size effects than experimental studies.[58]

[57] Of course, one explanation for these differences in findings is that class size effects are likely context-specific. However, this cannot explain why studies from the same country that cover the same grades come to very different conclusions (see e.g. Hoxby, 2000; Krueger, 1999).

[58] Whenever we talk about negative class size effects we mean worse student outcomes in larger classes.

Part one of this paper presents a model of a school system with two key features: (i) a grade retention rule by which students with academic skills below a certain threshold are redshirted (i.e. enrolled late in primary school) or retained, and (ii) exogenous shocks to the size of birth cohorts that translate into class size differences. The model delivers two main empirical predictions: First, within schools, the initial birth cohort size is negatively related to the grade-level share of students who have been held back in the past. Intuitively, in larger cohorts retained students from the previous (smaller) cohort mechanically make up a smaller share of students in the given (larger) cohort. Second, this negative association leads to a positive bias in class size estimates based on within-school variation in initial birth cohort size. This bias arises because larger cohorts experience, on average, larger classes but a lower share of negatively selected students — those retained from the previous cohort —, which increases average test scores in these classes. Since grade retention or delayed enrollment in primary school is a common practice in most countries,[59] our theoretical results have important implications for the majority of studies based on the within-school design.

We further propose a simple solution to this problem that is motivated by the following observation. The source of the upward bias is the negative relationship between cohort size and the share of negatively-selected students in higher grades. Simply adjusting the test scores of those negatively-selected students eliminates this link and produces estimates free of the resulting bias. Correcting can, therefore, be achieved by simply controlling for whether or not a student has previously been held back a grade.

In part two, we test our model's main predictions empirically using administrative school-level and student-level data from the German state of Saarland. In line with the first prediction, we show that birth cohort size is systematically related to the composition of students at the grade-level. Students from larger cohorts are enrolled in classes with a significantly smaller share of students who have been redshirted or retained in the past. Importantly, we can show that these compositional effects do not exist at the birth cohort level, i.e. students who are born into larger birth cohorts are not more or less likely to

[59] For example, the United States and 88 percent of European Union countries permit grade retention starting in primary school (European Commission, 2011).

be enrolled late. This is consistent with a purely mechanical effect driving the observed relationship between initial birth cohort size and student composition at the grade-level.

Our empirical results allow us to quantify the expected bias in class size estimates from within-school research designs that rely on birth cohort variation. The results imply that the bias can be expected to decrease estimates of a 10-student-reduction in class size between grades 1 to 3 on test scores in grade 3 by about 7.4 to 9.4 percent of a standard deviation. The magnitude of this bias is considerable and can be shown to increase even further in settings with higher retention rates or when test scores in higher grades are used as outcome variables. Since the share of retained students in German primary schools is at 7.7 percent similar to the OECD average of 7 percent (OECD, 2011; Ikeda and Garcia, 2014), we expect our results to be generalizable to countries that practice grade retention or delayed enrollment.[60] This insight recommends caution in the application and interpretation of within-school designs based on idiosyncratic variation in cohort size in school systems that allow for redshirting or grade retention.

Based on these considerations, we estimate class size effects utilizing data that cover four full cohorts of students in Saarland who participated in state-wide centralized exams in language and math at the end of grade 3 merged with administrative data on enrollment in grade 1. As an instrument for class size in grade 3, we use within-school variation in predicted class size based on changes in initial cohort size. In line with our theoretical model, adding a proxy for whether or not a student has been redshirted or retained in the past, leads to a substantial increase in effect size. Overall, we find that a one-student decrease in class size in grades 1 to 3 improves language and math test scores at the end of grade 3 by around 1.9 and 1.4 percent of a standard deviation, respectively. We interpret these estimates as lower bounds on the true effect sizes. Our study provides the first causal evidence of significant class size effects on test scores in Germany.[61] The

[60] Unfortunately, official statistics on delayed primary school enrollment are not available for most countries.

[61] Previous quasi-experimental studies for Germany cannot conclude that smaller classes improve student achievement. Wößmann (2005) is the only study that analyzes the effect of class size on test scores but the standard errors are too large to be able to detect our average effects at the 95 percent level of statistical confidence. Argaw and Puhani (2018) study the relationship between class size and recommendations for track choice in secondary school and actual track attendance as well as grade repetitions in another German state (Hesse). They find no or small effects on tracking, but a higher likelihood of repeating a grade in larger

beneficial impact of smaller classes is also supported by our finding that retention rates drop by 0.15 percentage points (7 percent) if the number of students in a class is reduced by one.

However, these average effects mask a significant degree of heterogeneity. We find class size effects to be non-linear, with large effects in larger and no effects in smaller classes. A one-student reduction in size in classes with more than 20.5 students (which is close to the average class size in our data) is predicted to improve language and math test scores by 4.8 and 3.8 percent of a standard deviation. At the same time, we uncover no evidence that class size reductions improve student outcomes in classes smaller than 20.5 students. Moreover, in line with Krueger (1999) our results suggest that disadvantaged students benefit the most from attending smaller classes: for example, the test scores of students with insufficient German proficiency or a learning disability are predicted to increase, on average, by around 3.5 to 4.1 percent of a standard deviation in language and 2.4 to 4.4 percent of a standard deviation in math for a one-student decrease in class size. Overall, these effects are large and similar in magnitude to those from the randomized experiment Project STAR.

These heterogeneous patterns have important policy implications. The larger benefits of smaller classes for disadvantaged children warrant the use of progressive maximum class size rules. These rules prescribe smaller maximum class sizes as the number of disadvantaged children in a grade increases. Saarland is one of several German states that practices these flexible rules. Furthermore, class size reductions to increase student achievement only seem to be efficacious in larger classes. Hence, class size reductions should be targeted at larger classes. Indeed, the finding of no beneficial effects of smaller classes in small classes, indicates that class size may be increased up to a certain size without negative consequences for student achievement.

Going back to our theoretical results, we expect that our simple solution to correct for the upward bias in within-school estimates provides an opportunity for researchers to revisit this empirical strategy to further investigate class size effects in other contexts. This

classes.

is important since within-school designs provide a number of advantages over commonly applied “Maimonides”-style research designs that exploit variation in class size generated by maximum class size rules as pioneered by Angrist and Lavy (1999) and subsequently used in numerous other studies.[62] First, the within-school design is widely applicable and allows for studying class size effects even if no class size rules exist or when the correct class size threshold cannot easily be identified, because different thresholds are in place that depend on characteristics unobservable to the researcher.[63] Second, regression discontinuity designs (RDD) can yield biased estimates in some contexts where carefully implemented within-school designs may not.[64] Gilraine (2018), for example, shows that crossing the class size threshold in New York City often prompts the hiring of a teacher of below-average quality. The resulting discontinuity in teacher quality substantially biases RDD class size estimates upwards. Moreover, our finding that grade retention rates increase with class size could result in a discontinuous change in the student composition at the class size threshold, which is also likely to bias RDD estimates of class size effects. Third, within-school designs allow the estimation of heterogeneous class size effects along the full range of the class size distribution. The advantage of this flexibility is the ability to detect the type of non-linear effects that we find in our data, which are missed in RDDs.

The rest of the chapter is organized as follows: Section 4.2 discusses the related literature. Section 4.3 develops our theoretical model and its implications for previously used research designs. Section 4.4 sets out the institutional background for our empirical part. Section 4.5 presents our estimation strategy. Section 4.6 describes the data used in our analysis. Estimates are presented and interpreted in section 4.7, with conclusions drawn in section 4.8.

[62] This regression discontinuity approach is used to study the effects of class size by Hoxby (2000) in the United States, Dobbelsteen et al. (2002) in the Netherlands, Browning and Heinesen (2007), Krassel and Heinesen (2014) and Nandrup (2016) in Denmark, Bressoux et al. (2009) and Piketty and Valdenaire (2006) in France, Asadullah (2005) in Bangladesh, Wößmann (2005) in 10 European countries, Jakubowski and Sakowski (2006) in Poland, Urquiola (2006) in Bolivia, Angrist et al. (2017a) in Italy, Falch et al. (2017) and Leuven and Oosterbeek (2018) in Norway, and Argaw and Puhani (2018) in Germany.

[63] In our empirical application, for example, the class size threshold depends on the number of students with insufficient German proficiency in first grade. Since we have no information on students' German proficiency in first grade, we cannot assign the correct class size thresholds.

[64] See e.g. Urquiola and Verhoogen (2009); Cohen-Zada et al. (2013); Gilraine (2018).

4.2 Literature Review

While the study of class size effects dates back at least to the early 1920s (Stevenson, 1922), we will focus here on more recent experimental- and quasi-experimental attempts to identify causal class size effects.[65] The methods applied in these studies can be broadly classified into three categories. The first is randomized experiments. Tennessee's Student Teacher Achievement Ratio Project — "Project STAR," as it is known — is the largest and most influential class size experiment ever conducted. Primary school students were randomly assigned to classes of different sizes during kindergarten and the first three years of schooling. Krueger (1999) provides a careful analysis of this project and finds a significant negative effect of class size on achievement. Students assigned to small classes performed five to seven percentile points (0.20-0.28 SD) better than students assigned to regular classes, which had on average about seven students more. Project STAR seems to have had long-run effects reaching well into adolescence and young adulthood as shown by a higher likelihood of graduating from high school and college enrollment and higher labor market earnings (e.g. Krueger and Whitmore, 2001; Finn et al., 2005; Chetty et al., 2011). Molnar et al. (1999) provide more experimental evidence of class size effects by evaluating the Wisconsin SAGE program which was considerably smaller than Project STAR. They find class size effects of similar magnitude to those from Project STAR.

A second common strategy to identify class size effects, hereinafter referred to as the within-school design, was first introduced by Hoxby (2000). The underlying idea of this approach is to leverage variation in class size arising from random fluctuations in cohort size that occur within a particular school (or school district) over time to obtain causal class size estimates. Hoxby (2000) uses school-district-level data from Connecticut.[66] As an instrument for the average class size a cohort from a specific district has experienced up until the time of the test (which is either in 4th or 6th grade), Hoxby uses the number

[65] Rockoff (2009) reviews the early pre-1940 literature. See Hanushek (1986, 1989, 1996, 1998) for summaries of the literature from the 1950s to the 1990s and Krueger (2003) for a reassessment of that literature.

[66] Using school-district instead of school-level data allows to rule out biases resulting from time-variant-selection of students into different schools within a school district, with the limitation that the identifying variation is substantially reduced.

of five-year-old children in each school district from the year that a particular cohort should have been enrolled in kindergarten according to the school entry rule.[67] To isolate natural randomness in birth cohort sizes from any secular trends, she controls for flexible school-district trends using 24 years of birth cohort data.[68] Her results indicate no class size effects and rule out effect sizes as small as 0.04 SD for a 10 percent reduction in class size.[69] The same approach has been used to study class size effects in Norway and Minnesota by Leuven et al. (2008) and Cho et al. (2012), respectively. While Cho et al. (2012) find small significant effects, Leuven et al. (2008) find no effects.

The type of data required for this approach, namely a long panel of demographic data merged with test scores data, are often not available to researchers. Instead, many studies use slight variants of Hoxby's approach and regress student test scores directly on the school's average class size in the grade at the time of the test while controlling for school fixed effects.[70] We have listed all within-school studies that we could find and broken them down along a number of dimensions in Table A4.1. All studies use data from school systems that allow either for grade retention or redshirting of students.[71] While differences in grades covered, the aggregation level of data, and other factors cloud comparisons of the magnitude of class size effects across these studies, none of the listed within-school design studies find effect sizes as large as those from Project STAR.[72] In fact, of the 11 papers summarized, four find no significant class size effects and one even finds signif-

[67] The school cohort here refers to the group of students who are in the same grade at the time of the test. These are not necessarily students from the same birth cohort if the school system allows for grade retention or the late enrollment of students, which is the main reason why this instrumental variable strategy could lead to biased estimates, as will be discussed below.

[68] Hoxby is also careful to distinguish between cases where the population variation triggers the opening or closing of a class (through a maximum class-size rule), and where it only causes variation in class size without opening or closing a class. This can be achieved by including fixed effects for each school/expected-number-of-classes combination.

[69] Hoxby (2000) uses the natural log of class size as an explanatory variable. Hence, her estimates measure the effect of a proportionate change in class size.

[70] Some studies instrument actual class size with the average class size in that grade and year if the data do not include all classes from a school in a given grade.

[71] However, not all school systems in these analyses allow for both redshirting and grade retention. Denny and Oppedisano (2013), for example, investigate class size effects with PISA data from the United States and the United Kingdom. Whereas grade retention and redshirting is very rare in the United Kingdom, it is relatively common in the United States.

[72] As is well known, effect sizes tend to be inflated with the level of aggregation. For example, effects sizes with school-district-level data are measured in the standard deviation of test scores by school-district-year, which is, of course, smaller than the standard deviation of individual student test scores.

icant beneficial effects of larger classes. The main identifying assumption under which estimates of these studies have a causal interpretation is that the within-school variation in cohort size is not related to any determinants of student achievement other than class size. However, even if this assumption holds true, class size estimates may suffer from a bias if the school system allows for academically weak students to be held back.

The third popular strategy to identify class size effects exploits maximum class size rules in a regression discontinuity design. This approach was first used by Angrist and Lavy (1999) and Hoxby (2000) and has since been applied in various studies spanning many countries. Gilraine (2018) and Leuven and Oosterbeek (2018) provide summaries of those papers. Gilraine (2018) reports that only three out of the 14 papers he summarizes find effect sizes qualitatively similar to those from Project STAR. The majority of papers cannot conclude that class size affects student achievement. As some studies have pointed out, however, depending on the institutional context, RDD estimates of class size effects may be prone to substantial biases. Bias may be introduced if school principals are able to manipulate enrollment around the maximum class size cutoffs or if crossing a cutoff leads to the hiring of a lower quality teacher (Urquiola and Verhoogen, 2009; Cohen-Zada et al., 2013; Gilraine, 2018). Our paper points out yet another potential source of bias that arises if class size affects retention rates and thereby the composition of classes with enrollment just below and above the maximum class size cutoffs. These findings cast doubt on the validity of the identifying assumptions in some of the RDD studies on class size effects.

4.3 Theoretical Model and Implications

4.3.1 Model of a School System with Grade Retention

To examine the validity of within-school designs to estimate class size effects, we extend the model of a school system with grade retention proposed by Ciccone and Garcia-Fontes (2015) below.[73] Our model differs in that it accommodates classes of different sizes, thus allowing to study how shocks that translate into differences in class size affect observed

[73] Naturally, this section draws heavily on Ciccone and Garcia-Fontes (2015).

test scores in higher grades.[74] This helps to clarify what parameters are identified in different empirical designs.

In each year t a new cohort that consists of a continuum of students with mass N_s^t starts primary school in school s. To simplify the model, we assume that schools have only one class per grade, such that the number of students per grade and school corresponds to actual class size.[75] Our model consists of two phases. We assume that students spend the first L school years in lower grades (LG). At the end of the Lth year in primary school, students move to higher grade (HG) if their academic skills a are higher than their school's academic threshold for grade retention p, i.e.

$$a_{is}^t > p_s^t \tag{4.1}$$

where a_{is}^t is the academic ability of student i in school s from cohort t and p_s^t is the retention threshold for school s and cohort t. Students with skills below the academic threshold $a_{is}^t < p_s^t$ spend another year in LG and move to HG after $L + 1$ years in LG.[76] We assume that the size and the grade retention threshold of cohorts are distributed with school-specific means

$$N_s^t = N_s + \eta_s^t \tag{4.2}$$

$$p_s^t = p_s + \nu_s^t \tag{4.3}$$

where η_s^t and ν_s^t are i.i.d. shocks at the school-year level with mean zero and positive variance (i.e. $Var(\eta_s^t) > 0$ and $Var(\nu_s^t) > 0$).[77] The distribution of individual students' skills in cohort t in school s after L years in LG, a_{is}^t, is taken to be uniform with density

[74] Ciccone and Garcia-Fontes (2015) set up a model that allows to study the effects of the gender composition of birth cohorts on the skills of students. Class size is kept constant in their model.

[75] Hence, we abstract from maximum class size rules that determine the number of classes per grade, but our view is that accounting for these rules would add more tedious complications than real insight. However, in simulations, which we do not report here, we can show that the implications of our model for the estimation of class size effects also hold if there are more than two classes in a school-year cell. We return to this issue in section 4.7.1.

[76] We assume that students can be retained only once.

[77] If the assumption of i.i.d. shocks to the size of birth cohorts is relaxed to allow for serial autocorrelation in η_s^t, it can be shown that under certain conditions, the positive bias to be derived below is increased. We explore this extension in Appendix D.

$1/2\theta$ and a school-cohort specific mean α_s^t. To capture class size effects in LG, the school-cohort specific mean in accumulated skills depends on class size in LG as follows

$$\alpha_s^t = \alpha_s + \pi^\alpha N_s^t + \epsilon_s^t \tag{4.4}$$

where π^α is the effect of class size in LG on academic skills and ϵ_s^t are i.i.d. shocks with mean zero and positive variance. In combination with the rule for grade retention in Equation (4.1), this implies that the share of students (λ) in cohort t who are not retained and hence reach HG in year $t+L$ is[78]

$$\lambda_s^t = \frac{\alpha_s^t + \theta - p_s^t}{2\theta} \tag{4.6}$$

Class size in HG in school s in the school year starting in τ depends on the size of cohort $\tau - L$ and the share of non-retained students in that cohort as well as the size of cohort $\tau - L - 1$ and the share of retained students in that cohort

$$N_{s\tau}^{obs} = \lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L-1}) N_s^{\tau-L-1} \tag{4.7}$$

The share of non-retained students in HG in school s in the school year starting in τ is therefore

$$\phi_s^\tau = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{N_{s\tau}^{obs}} = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{\lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L-1}) N_s^{\tau-L-1}} \tag{4.8}$$

In HG students acquire skills equal to $w_{is\tau}$, which are obtained as i.i.d. draws from a distribution with constant variance and a school-cohort specific mean $\omega_{s\tau}$ that is a function of class size in HG

$$\omega_{s\tau} = \tilde{\omega}_{s\tau} + \pi^\omega N_{s\tau}^{obs} \tag{4.9}$$

where π^ω captures the effect of class size in HG and $\tilde{\omega}_{s\tau}$ are exogenous shocks. Thus, the sum $\pi^\alpha + \pi^\omega$ captures the combined effect of class size in LG and HG on accumulated

[78] To ensure that the share of students who are not retained in LG in each school is between zero and one, we impose the following parameter restriction:

$$-\theta \leq \alpha_s^t - p_s^t \leq \theta \tag{4.5}$$

academic skills. This is our main parameter of interest, which we will refer to as the "pure class size effect." At the end of HG, students take a standardized test. The average test performance of non-retained students reflects their academic skills accumulated in LG and HG, $a_{is}^{t}+\omega_{is,t+L}$. The average test performance of these students from cohort t who reach HG in year $\tau=t+L$ can be written as

$$E\left(test_{is}^{t}|non-retained\right)=E\left(test_{is}^{t}|a_{is}^{t}\geq p_{s}^{t}\right)=\frac{\alpha_{s}^{t}+\theta+p_{s}^{t}}{2}+\omega_{s,t+L} \tag{4.10}$$

where $E\left(a|a\geq p\right)$ denotes the average skills of non-retained students in HG and $\omega_{s,t+L}$ denotes the average skills these students accumulate in HG in year $t+L$. The test performance of retained students who reach HG one year later is $a_{is}^{t}+w_{is,t+L+1}+\delta_{s}^{t}$, where δ_{s}^{t} captures a school and birth cohort specific change in skills associated with grade repetition. This change in skills may be positive or negative. The average performance of these retained students in HG is

$$E\left(test_{is}^{t}|retained\right)=E\left(test_{is}^{t}|a_{is}^{t}<p_{s}^{t}\right)=\frac{\alpha_{s}^{t}-\theta+p_{s}^{t}}{2}+\delta_{s}^{t}+\omega_{s,t+L+1} \tag{4.11}$$

where $E\left(a|a<p\right)$ denotes the average skills in HG of students who were retained. The average test performance of all students in HG in year τ can be derived by combining (4.8), (4.10) and (4.11)

$$test_{s\tau}=\phi_{s}^{\tau-L}E\left(test_{is}^{\tau-L}|non-retained\right)+(1-\phi_{s}^{\tau-L})E\left(test_{is}^{\tau-L-1}|retained\right) \tag{4.12}$$

So far, we have only modeled grade retention between LG and HG in primary school. However, it is straightforward to modify this framework to either capture redshirting (i.e. keeping students another year in child care before enrolling in primary school) or the early enrollment of children with accelerated maturity. This is important as redshirting and early enrollment have similar implications for the estimation of class size effects as grade retention. To model these differences in the timing of school enrollment, LG would refer to the last year in child care before primary school entry and HG would refer to the first grade of primary school. Children are redshirted if their skills fall below a certain threshold. Similarly, students with skills above a higher threshold enter HG one year earlier than planned. These models are explored more fully in Appendix C.

4.3.2 Model Implications

A useful starting point to understand what is identified through different within-school empirical designs in school systems of the type modeled in the previous section is the special case that resembles experimental conditions. In this setting, where everything is assumed to be constant across schools and cohorts and only initial cohort size is randomly assigned, it can be shown that commonly used within-school empirical designs are unable to identify the pure class size effect.[79] The main reason is that within-school differences in initial cohort size are positively correlated with within-school differences in test scores in HG. The easiest way to see this is by assuming that there is no pure class size effect (i.e. $\pi^{\alpha} = \pi^{\omega} = 0$). The instrumental variable approach exploiting variation in cohort sizes amounts to dividing the covariance of within-school changes of test scores in HG and within-school changes in cohort size by the covariance of within-school changes of cohort size in HG and initial cohort size. In Appendix D, we show that if there are no class size effects this ratio is equal to

$$\frac{3(\theta - \delta)(1 - \lambda)\lambda}{3\lambda - 1} \tag{4.13}$$

where $(\theta - \delta)$ is the average test score difference of non-retained students and students retained in the past, see Equation (4.10) and Equation (4.11), while λ is the average fraction of students who are not retained in LG. If $(\theta - \delta)$ is positive, i.e. non-retained students have higher skills, on average, than students retained in the past, it is easy to see that using the initial cohort size as an instrument will yield a spurious positive effect of class size if more than one-third of students are not retained in LG ($\lambda > 1/3$).

To develop some intuition for this result, consider the following thought experiment. Imagine a school that is in equilibrium but experiences a positive shock $\eta_s^t > 0$ to the size of cohort t, N_s^t. We show that this positive shock translates into changes in the size

[79] In the experimental setting $N_s = N, \alpha_s^t = \alpha$, $p_s^t = p$, $w_s^t = w$ and $\delta_s^t = \delta$. This also implies that $\lambda_s^t = \lambda$. The only shocks are shocks to initial class size η_s^t, as modeled in Equation (4.2).

of classes in HG as well as into changes in the share of retained students in HG, which results in the spurious effect in (4.13). First note that this shock increases the number of students from cohort t reaching HG after L years without being retained in LG by $\lambda\eta_s^t$. Therefore, cohort size in HG in year $t+L$ increases by $\lambda\eta_s^t$ from year $t+L-1$.[80] At the same time, the number of students who are retained in LG and reach HG in year $t+L+1$ is increased by $(1-\lambda)\eta_s^t$. Relative to year $t+L+1$, this implies an increase in class size in HG in year $t+L$ of $(2\lambda-1)$ for each additional student in cohort t. Hence, it depends on the share of retained students whether the association between a positive shock to cohort size in year t and the change in class size in HG between the years $t+L$ and $t+L+1$ is positive or not. However, as long as less than half of all students are retained, this association will be positive.

In brief, a positive shock to the size of cohort t leads to a positive association between the difference in initial cohort size between cohort t and $t-1$ and class size in HG L years later and, if less than half of all students are retained, also a positive association between the difference in initial cohort size between cohort t and $t+1$ and class size in HG L years later. The covariance of within-school changes in class size in HG and initial cohort size ends up summing up these two associations, λ and $(2\lambda-1)$, which explains the denominator in (4.13).[81] Therefore, the sign of the first stage in an instrumental variable approach where class size in HG is instrumented with initial cohort size will generally be positive if less than two-thirds of all students are retained.

Crucially, the positive shock to cohort size in year t also translates into within-school changes in the composition of students in HG, and, therefore, a positive reduced form coefficient. To see this, note that retained students from cohort $t-1$, who join HG in year $t+L$, will account for a smaller share of students in that grade compared to year $t+L-1$. This is because the number of non-retained student in year $t+L$ increases by $\lambda\eta_s^t$ as a result of the positive cohort shock in year t, while the number of retained students who join HG in year $t+L$ remains constant. At the same time, the additional students from cohort t who were retained $(1-\lambda)\eta_s^t$ will increase the share of retained students in year

[80] Recall that cohort size in HG in year $t+L-1$ is equal to the equilibrium value N.

[81] It is easy to see that the covariance of first differences in class size in HG and initial class size is equal to $Var(\eta)(3\lambda-1)$. However, $Var(\eta)$ cancels out in (4.13) because it also appears in the numerator.

$t + L + 1$ and, therefore, further decrease the relative share of retained students in year $t + L$ compared to $t + L + 1$.

Together, these two effects imply that a positive shock to cohort size in year t will always be associated with a reduction in the share of retained students in HG in year $t + L$ relative to $t + L - 1$ and $t + L + 1$. If non-retained students have, on average, higher skills than retained students in HG, test scores will be higher in $t + L$ than in $t + L - 1$ and $t + L + 1$. In turn, this translates into a positive reduced from coefficient in a within-school regression of test scores in HG on initial cohort size. This spurious effect is central for the understanding of what parameters are identified by different research designs. In instrumental variable terminology, using initial cohort size as an instrument to identify the effect of class size on student achievement leads to a violation of the exclusion restriction due to the share of retained students at the grade-level being negatively correlated with the instrument even if initial cohort size is random. Since the first-stage has a positive sign if $\lambda > 1/3$, this results in a positive spurious effect of class size on test scores. Ciccone and Garcia-Fontes (2015) identify a similar bias in the analysis of gender peer effects where shocks to initial gender composition of cohorts also translate into positive peer effects even in the absence of true peer effects.

Analogous arguments show that, in a school system that allows for redshirting or early school enrollment, there will be similar spurious class size effects, the sign of which depends on whether redshirted or early enrolled students have, on average, lower or higher skills than students who reach HG on schedule.

4.3.2.1 Instrumental Variable Approach

Using this setup and the previous result, one can clarify the parameters identified in an instrumental variable approach exploiting birth cohort variation. Suppose we observe the test performance and class size in HG as well as the class size students should have started out with if they were not retained for all students from a large number of schools for two consecutive years (i.e. we observe $\{N^{obs}_{s\tau}, N^{obs}_{s,\tau-1}, test_{s\tau}, test_{s,\tau-1}, N^{\tau-L}_{s}, N^{\tau-L-1}_{s}\}$).[82]

[82] It would be straightforward to extend our results to a setting with data for more than two years. But this would not generate further insights as far as we can see.

The commonly used instrumental variable approach would estimate class size effects by regressing individual test performance in HG for year τ on school fixed effects and class size in HG for year τ while instrumenting class size in HG by the respective cohort size in year $\tau - L$.[83] In Appendix D, we show that in this setup, where shocks to the initial cohort size are completely independent from shocks to the academic skills and shocks to the grade retention thresholds, the IV estimate will converge in probability to

$$\beta_{IV} = \underbrace{(\theta - \delta)\rho_{IV}}_{\text{grade retention bias I}} + \underbrace{\xi_{IV}}_{\text{attenuation factor}} \pi^{\alpha} + \pi^{\omega} \tag{4.14}$$

where ρ_{IV} is a function of λ and $\pi^{\alpha}/2\theta$ that takes on strictly non-negative values for a wide range of plausible values for these parameters.[84] If students previously retained have lower average academic skills than non-retained students (as in our data), this will cause a positive bias in the IV estimate of class size effect in HG. This bias is a result of the positive correlation between initial cohort size and the share of non-retained students in HG as discussed above.[85]

ξ_{IV} is a function of λ, $\pi^{\alpha}/2\theta$, $Var(\epsilon)$, and $Var(\nu)$ and can be shown to only take on values well below one, which implies an attenuation bias for the class size effect in LG, π^{α}. This is similar to the standard classical attenuation bias because our explanatory variable class size in HG is a noisy measure of class size in LG for two reasons: First, class size in HG is not perfectly correlated with class size in LG because retained students lead to changes in the size of the same class between these grades. Second, the observed class size in HG for students who were retained in LG should be at most weakly correlated with the class size these students experienced in LG.[86] The importance of this attenuation

[83] Most studies do not directly use cohort size as an instrument. Instead, they regress cohort size on higher polynomials of time separately for each school catchment area (or school district). The residuals from these regressions are then used as an instrument for class size. Thereby, differences in cohort size stemming from smooth variations over time are removed. Our findings carry over to these approaches. Additionally, the number of classes is held constant so that increases in cohort size are always associated with larger classes. This ensures that the monotonicity assumption of the instrumental variable is not violated.

[84] See Appendix D for more details.

[85] Unlike expression (4.13), ρ_{IV} does not just depend on λ but also on $\pi^{\alpha}/2\theta$. The reason is that the initial cohort size, N_s^t, affects the retention rate in LG ,$1 - \lambda_s^t$, if $\pi^{\alpha} \neq 0$; therefore also $test_{s,t+L}$ and $test_{s,t+L+1}$. However, this should have a negligible impact on the size of the bias, as shown in Appendix D.

[86] Although we do not model this explicitly, it is easy to see that students switching schools will exacer-

bias has previously been pointed out by Jepsen and Rivkin (2009).

These two sources of bias imply that even if initial cohort size is unrelated to academic skills and grade retention thresholds, the net effect of the bias will likely be upwards, i.e. reduce the estimated size of the negative class size effect. In the appendix, we further show that this bias increases with the retention rate, $1 - \lambda$. A natural solution for the first bias is to control for the effect of grade retention on academic achievement at the individual level.[87] In the appendix, we prove that by conditioning on whether a student has been retained the IV estimator will consistently estimate

$$\beta_{IV}^{REA} = \xi_{IV}\pi^{\alpha} + \pi^{\omega} \tag{4.15}$$

where REA stands for retention-effect adjusted. To get an intuition for this result, recall that the bias $\rho_{IV}(\theta - \delta)$ is a result of the positive correlation between cohort size and the share of non-retained students in HG. Since non-retained students have higher average academic skills than retained students, this translates into a positive correlation between initial cohort size and test scores in HG. However, conditioning on grade retention removes any correlations in test scores that are solely driven by differences in the share of retained students as long as the difference in skills between retained and non-retained students is not correlated with shocks to initial cohort size. So while conditioning on grade retention removes the positive grade retention bias, it does not resolve the attenuation of the class size effect in lower grades. The resulting estimate in Equation (4.15) thus yields a lower bound of the true class size effect.

4.3.2.2 OLS Approach

Instrumental variable estimates generally have large standard errors that reduce the power to detect class size effects. In addition, oftentimes it is not possible to match birth cohort

bate both sources of attenuation bias. Students switching schools will increase the differences in the size of the same class between lower and higher grades, thereby reducing the correlation between class size in LG and HG. At the same time, if students change schools and join a new class in HG, the size of that class is an erroneous measure of class size in their previous class at a different school.

[87] Ciccone and Garcia-Fontes (2015) show a similar result for the case of peer effects contaminated by grade retention.

size information to student test score data. Many studies in Table A4.1, therefore, regress test scores directly on observed class size in HG conditional on school fixed effects since this places a substantially lower demand on the data relative to the IV approach. In Appendix D we show that in our set-up the resulting estimate will converge to

$$\hat{\beta}_{OLS} = \underbrace{(\theta - \delta)\rho_{OLS}}_{\text{grade retention bias I}} + \underbrace{\iota_{OLS}}_{\text{grade retention bias II}} + \underbrace{\xi_{OLS}}_{\text{attenuation factor}} \pi^{\alpha} + \pi^{\omega} \tag{4.16}$$

Here we have three sources of bias. The first bias, $(\theta - \delta)\rho_{OLS}$, results from the correlation between class size in HG and the share of grade repeaters in HG, which is similar to the instrumental variable result in Equation (4.14). ρ_{OLS} differs slightly from its IV counterpart, but it can still be shown to take on strictly positive values. The source of the second bias, ι_{OLS}, are shocks to ability levels and grade retention thresholds that lead to differences in class size in HG as well as to differences in skill levels between retained and non-retained students in HG.[88] The sign of ι_{OLS} depends on the relative magnitude of these shocks. Since they are unobserved, it is impossible to tell what the net effect of the bias on $\hat{\beta}_{OLS}$ will be. However, comparing IV and OLS estimates could give us a sense of the direction and magnitude of this bias. The third bias is again caused by measurement error as class size in HG is not perfectly correlated with class size in LG. The attenuation factor ξ_{OLS} for the class size effect in LG also differs slightly from its IV counterpart, but can still be shown to take on values strictly below one.

Analogous to the IV case, controlling for grade retention at the individual level removes the first bias

$$\hat{\beta}_{OLS}^{REA} = \iota_{OLS} + \xi_{OLS}\pi^{\alpha} + \pi^{\omega} \tag{4.17}$$

However, ι_{OLS} does not disappear because it is the result of shocks that cause ability levels of retained and non-retained students to deviate from their respective average values. Moreover, estimates will still be attenuated. Albeit more susceptible to bias, this

[88] Shocks to student ability, ϵ_s^t, and retention thresholds, ν_s^t, can be shown to lead to differences in average test score differences of non-retained and students retained in the past, $E(test_{is}^{\tau-L}|non-retained) - E(test_{is}^{\tau-L-1}|retained))$, which are correlated with $N_{s\tau}^{obs}$. IV estimates do not suffer from this second bias as long as these shocks are uncorrelated with shocks to the initial cohort size.

OLS estimator should be more efficient than the IV approach based on initial cohort size.

The above results are easily extended to school systems that allow for redshirting or early school enrollment. We explore these extensions more fully in Appendix C.

4.4 Institutional Context

To empirically investigate the implications of our model, we focus our empirical analysis on one German federal state (Saarland), for which we have detailed student test score data for multiple years of all third-graders. Generally, all federal states in Germany run their own educational systems, but states agree on some common standards so that many features are shared across states. This is especially true for primary education. As a result, most characteristics of primary schooling in Saarland are similar to all other German federal states. Primary school in Saarland is obligatory, free of charge and spans grades 1-4. School entry is determined by a cut-off date set at June 30th. Children turning six before this cut-off start school at the beginning of the same school year. Children born after the cut-off are enrolled in the next school year. However, children may be sent to school in the year before or after they become eligible depending on their maturity.[89] There is no explicit ability tracking in primary school.[90] Furthermore, it is not possible to fail one of the first two grades in Saarland. However, children may be retained in these grades with their parents' approval.

Allocation of children to primary schools is determined by place of residence with little choice for parents since primary schools have well-defined catchment areas that generally do not overlap. Only a handful of all-day schools have catchment areas that overlap with

[89] Early school entry is possible upon parental request subject to the school principal's agreement. Principals base their assessment on the results of a medical- and in some cases a psychological examination of the child as well as a talk with the parents. Equally, principals may decide to defer school entry for another year. For this to happen, a number of requirements must be fulfilled. First, the results of the obligatory diagnostic language tests in the year before regular school entry have to be unsatisfactory. As a result, parents would usually be advised to send their child to a special preparatory course in the following year. Only if this course does not bring about the desired improvement or if parents fail to follow the advice altogether, principals may reject applications for regular school entry (Lisker, 2010).

[90] While Germany is known for early ability tracking, this happens only when students leave primary school after fourth grade and enroll at one of three different secondary schooling tracks (Gymnasium, Realschule or Hauptschule).

those of other schools (Ministerium für Bildung und Kultur, 2018). However, parents who are not satisfied with their assigned school have two options to change schools. First, they may send their child to a private school. In practice, however, very few parents resort to this option: private primary schools are rare in Germany. In 2006, there were only 624 of these schools which accounted for 3.7 percent of all primary schools in Germany (Autorengruppe Bildungsberichterstattung, 2016). Almost all of these schools were boarding schools, religious schools or schools offering specialized pedagogic approaches, like Waldorf education (Cortina et al., 2008). The second option, sending the child to a different public school, is only possible under certain conditions; for example, if a different school offers full-day care while the local school does not. Reasons pertaining to comfort or preference alone are generally not deemed sufficient to switch schools. Ultimately, school principals have to decide whether or not a claim is well-founded and, consequently, if the change of school should be granted. When making this decision, they are obliged to apply strict standards (Schulordnungsgesetz, 2006).

Like most countries, school funding in Saarland is a function of the number of classes in a grade. This number is determined by maximum class size rules. Prior to the 2002-03 school year, the maximum class size was set at 27 students (for ease of discussion we subsequently refer to an academic year by the calendar year in which it begins). Hence, whenever a class would exceed 27 students, a new class had to be formed. This threshold increased to 29 in the summer of 2003. However, if the average number of students with insufficient German proficiency per class was at least 4 in a grade, the threshold was set at 25 (Ernst, 2017). Note that class size is a much more meaningful concept in German primary schools than in secondary schools. Students are taught in the same classroom with the same peers in all or almost all subjects and the teacher is also the same in most subjects (Jonen and Eckhardt, 2006). The majority of students in a classroom stay together for the entire duration of primary school. Classroom composition changes only if children repeat grades, switch schools, or, in rare cases are moved to a different classroom of the same grade.

Importantly, during the school periods for which we have test data, Saarland enacted a major structural reform in the primary school sector. Due to decreases in the number of school-aged children, which drove up per-student costs especially in rural areas with low population densities, policy-makers decided to merge schools to ensure that all schools would have at least two classes per grade. This meant that primary schools with an insufficient number of students to form at least two classes per grade were merged with other primary schools. This applied to around one third of all schools. Hence, the number of primary schools decreased from 268 in 2004 to 159 in 2005. However, the reform was not practically implemented at once in all schools. In most places, almost all incumbent students continued to be taught in the same buildings and classrooms as before. Only new incoming cohorts were sent to the main building of the newly merged schools. Because even the most recent cohort for which we have test score data was already enrolled in primary school when this policy was enacted, the consolidation of schools had no discernible impact on the third graders in our data. Therefore, we do not exploit this policy reform for identification of class size effects. However, by estimating separate school fixed effects before and after consolidation for schools that were merged, we make sure that the reform does not bias our estimates.

4.5 Estimation Strategy

The main difficulties in the identification of class size effects arise from student sorting at various institutional levels. Parents self-select into neighborhoods and, within schools, students may be assigned to different classes of different sizes depending on their abilities. As students are typically not assigned to schools at random, studies using the within-school design try to overcome this identification issue by exploiting natural variation in cohort size within a given school across time. We follow this approach by estimating equations of the following form:

$$y_{icts} = \alpha_0 + \alpha_1 CS_{ts} + \alpha_2 X_i + T_t + S_s + \epsilon_{icts} \tag{4.18}$$

where y_{icts} represents the standardized test score of student i in class c in year t in school s; CS_{ts} is the average class size in grade 3 in school s in year t; X_i is a vector of student i's characteristics (e.g., gender); T_t is a year fixed effect, and S_s is the school fixed effect. Hence, we control for between-school sorting by using school fixed effects.

To circumvent any problems resulting from the potential sorting of students and teachers within the same year and school into classes of different sizes, we use average class size in a given school, grade, and year rather than actual class size.

Similar to existing studies, we only want to exploit arguably random variation in the timing and number of births in a school catchment area. Thus, ideally, we would estimate Equation (4.18) via 2SLS using the predicted class size based on a school's birth cohort size as an instrument for class size in grade 3. Unfortunately, data on the number of births at the level of the school catchment area are not available in Germany, but we can impute cohort size using administrative school-level data on enrollment in grade 1. For a given school in grade 3 in year t, we do this by summing up the number of regularly enrolled students in grade 1 in year $t-2$, the number of late enrolled students from year $t-3$, and the number of early enrolled students from year $t-1$. Dividing this sum by the number of classes in grade 1 in year $t-2$ gives the predicted class size for grade 3 in year t, which we then use as an instrument for CS_{ts} in Equation (4.18).

As discussed in section 4.3, estimating class size effects this way will result in biased estimates since birth cohort size should be correlated with the grade-level composition of students. To overcome this bias, we need to control for whether a student has been retained, enrolled late, or enrolled early at the individual level (i.e. include dummies for each group of students in the vector X_i). Since our test score data only contain age in years at the time of the test, we use separate dummies for each age as proxies for each group of students.[91] This amounts to combining students who have been retained or enrolled late into one group because both types of students are older than 9 years on the day of the test. Thereby, we also incorrectly assign those students reaching third grade one year late but who were born between May and June to the group of students who reach 3rd grade on time (recall that the enrollment cutoff is the 30th of June and age is measured in May). Therefore, we expect to underestimate the size of the pure class size effect for

[91] Note that controlling for age linearly, as done in some previous studies (see e.g. Wößmann and West, 2006; Denny and Oppedisano, 2013), is not sufficient to correct for the upward bias. The reason is that the negative relationship between age and test scores, caused by negatively selected students who are too old for their grade, is offset by a positive effect of age on test scores for students who are on schedule (Black et al., 2011). Hence, controlling linearly for age does not correctly adjust test scores for retained and redshirted students.

two reasons. First, assigning some retained- or redshirted students to the group of non-retained students decreases the average test score of the group of 9 year old students in our data. Effectively, this implies that we underestimate the average test score difference of non-retained students and students too old for their grade, $\theta - \delta$. Since the bias in Equation (4.14), $\rho_{iv}(\theta - \delta)$, is a positive function of this difference, we expect an upward bias in estimates of the pure class size effect. Second, our estimations do not adjust test scores of those students who reach 3rd grade late but who are reported to be 9 years old in our data. Our model predicts that the grade-level share of these students (who should have below average test scores) will be lower in years associated with larger initial birth cohorts. This should also upward bias our estimates.[92]

The fact that different maximum class size rules apply depending on the number of students with insufficient German proficiency in grade 1 introduces a further bias in class size estimates based on Equation (4.18). Because even if the cohort size across years within the same school is completely random, random shocks to the number of students with insufficient German proficiency in a cohort lead to a spurious positive class size effect if these students score lower on standardized tests (as in our data).[93] To reduce this upward bias, we can include in the vector X_i a dummy variable indicating whether the teacher reported that the student has insufficient German proficiency in grade 3. This is only a proxy for insufficient German proficiency in grade 1 as some students become proficient in German until grade 3. Hence, we expect this to only partially correct for the positive bias.[94]

Around one-third of all primary schools in Saarland were merged in 2005. This consol-

[92] Similarly, students who were born between May and June and enrolled on time, will be incorrectly classified as having been enrolled too early. However, this should not have an effect on our estimates as discussed further below.

[93] To see this, consider two cohorts in the same school with 27 students. Suppose that all students are identical in terms of their academic skills except that the second cohort includes 4 students with limited German proficiency who have academic skills considerably lower than all other students. Due to these 4 students, the maximum class size threshold of 25 applies for the second cohort, while the threshold 27 applies for the first cohort. Hence, class size will be 27 and 18.6 for the first and second cohort, respectively. Since the average skill is lower in the second cohort, a simple within-school regression of test scores on class size would result in a spurious positive class size effect.

[94] German proficiency in grade 3 is, of course, potentially endogenous because it might be affected by class size. However, since class size can be expected to negatively affect German proficiency, controlling for it provides a lower bound on the true class size effect.

idation of schools is a potential threat to our identification strategy since school-specific factors, such as material resources and the composition of students, may have changed as a result. These time-varying changes are not picked up by school fixed effects. For this reason, we estimate separate fixed effects for schools that were eventually merged on the individual school-level for the academic years 2003-2004 (when they were not yet merged) and on the consolidated school-level for the academic years 2005-2006.[95]

As discussed in section 4.3, the key identifying assumption for the IV approach to identify the lower bound of the true class size effect in grades 1 through 3, β_{IV}^{REA}, in Equation (4.15) is that birth cohort size within school catchment areas is not correlated with shocks to the ability level of cohorts, ϵ_s^t, or the academic thresholds determining early and late school enrollment and grade retention, p_s^t. The most obvious violation of this assumption comes from potential self-sorting of families into specific school catchment areas that is not constant over time. To assess the credibility of our assumption, we conduct an extensive set of balancing checks in section 4.7.2 in which we test whether the composition of cohorts is systematically related to their size.

4.6 Data

4.6.1 State-wide Orientation Exams

We use a unique administrative dataset that contains information on the math and language skills for the full universe of four consecutive cohorts of third-graders in the German state of Saarland.[96] [97] The data were obtained via state-wide centralized exams at the end of grade 3 in the school years 2003 to 2006. Participation in these "State-wide Orientation Exams" (SOE) was obligatory for all schools and classes.[98] Testing was carried out

[95] For efficiency reasons, we would ideally estimate only one set of fixed effects at the individual school-level for schools that were merged in 2005 in which 3rd grade classes continued to be taught in their old schools. However, in our data we do not observe to which school classes belonged before consolidation. Hence, the need to aggregate everything to the consolidated school level for merged schools.

[96] If not stated otherwise, all information provided in this section is based on Paulus and Leidinger (2009).

[97] Students who were educated with "different aims" (zieldifferent) were exempt from the exams. Education with different aims is often applied for students with disabilities.

[98] The only exception was a school where teaching was conducted exclusively in French.

on three different days — two days for language and one day for math. If a student was not present on the day of testing, she was not allowed to take the exam later and her test score is, therefore, missing. We provide more information on these data in Appendix B.

Standardized assessments may suffer from bias introduced by intentional teacher manipulation in answer sheet transcription (see e.g. Angrist et al., 2017a). In our case, there is an incentive for teachers to manipulate test scores, since the results directly affect them. It was a specific objective of the SOE to compare achievement between different schools and even between classrooms within schools in order to detect successful approaches to teaching and learning. To prevent the most common forms of teacher cheating and shirking, particularly teaching to the test and biased grading, the designers of the exams established a number of safeguards. First, teachers had to keep the test material sealed until the day of testing. That way, specific preparation for the test was prevented. Second, and most crucially, teachers did not correct the exams themselves. Answer sheet transcription and grading was performed by a team of scorers who followed the provided grading rubrics. Therefore, score manipulation by teachers can be ruled out.

We link the 2003-2006 test score data to administrative records obtained from the Saarland statistical office. These administrative records include enrollment and number of classes for grades 1-3 for all schools in Saarland. Furthermore, for the 2000-2005 school years, these data contain information on the school-year-level on the number of students in grade 1 who were retained, who were enrolled one year late and who were enrolled one year early. This information is used to impute initial cohort size. Table A4.2 shows the structure of the Saarland data by academic year.

4.6.2 Sample Selection, Variables and Descriptive Statistics

The full SOE dataset comprises 39,014 student-year observations from 268 schools. We impose a set of restrictions on these data. First, we drop all schools for which we observe zero classes for some years. These are schools that formed multi-grade-classes because enrollment was too low to form separate classes for each grade. This restriction means that we exclude 10 schools (less than 4% of all schools). Next, in order to reduce measurement error, we exclude individual students if the teacher indicated that the student arrived too late to class that day to be able to complete the test. This restriction results in less than 0.2% of our initial data being dropped. Our final dataset includes 37,847 language and 36,845 math test scores from 38,415 students.

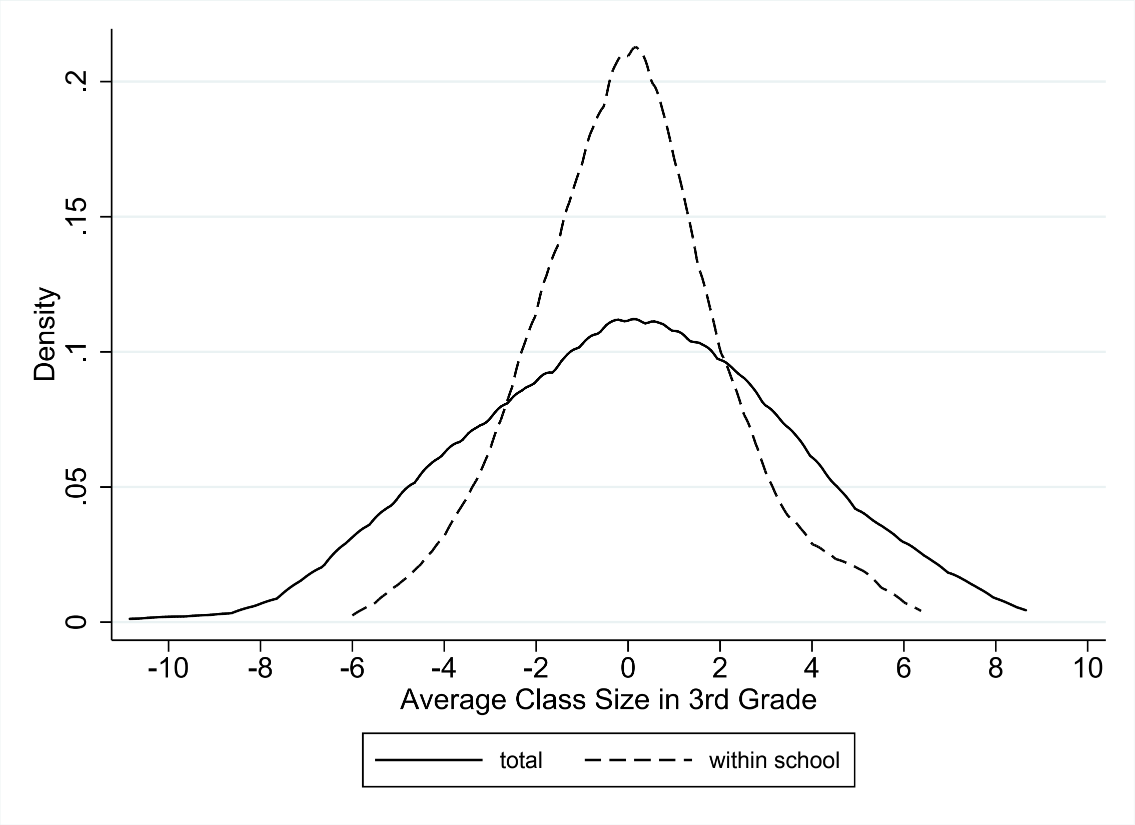
Table 4.1: Descriptive statistics: Student outcomes, student and school characteristics

	Mean	SD	N
Test scores			
Language	0.00	1.00	37,847
Math	0.00	1.00	36,845
Male	0.51	0.50	38,154
Insufficient German proficiency	0.06	0.23	38,415
Migration background	0.12	0.33	37,679
Non-native German speaker	0.15	0.35	37,920
Reported books at home			
None or few books	0.06	0.23	27,850
Enough to fill one shelf	0.17	0.37	27,850
Enough to fill one bookcase	0.26	0.44	27,850
Enough to fill two bookcases	0.26	0.44	27,850
$\geq$ 200 books	0.25	0.44	27,850
Age at test date (in years)			
Younger than 9	0.15	0.35	38,177
9	0.74	0.44	38,177
Older than 9	0.12	0.32	38,177
Learning disabilities			
Dyscalculia	0.04	0.19	37,314
Dyslexia	0.07	0.26	37,549
Class size grade 3	20.84	3.53	38,415
Cohort size	58.48	23.84	38,415
School district			
Rural community	0.54	0.50	38,415
Problematic	0.27	0.44	34,289
Classes per cohort	2.79	1.06	1,929
N Schools			258
N SchoolYearObs			828
N Cluster			156

Notes: The table reports means, standard deviations, and the number of non-missing observations for the listed variables. The sample only includes schools with at least one class for each grade. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table 4.1 reports descriptive statistics for our final sample. We standardize test scores to have mean zero and a SD of one. Note that we keep observations from students who participated in only one of the two days of testing in German. This applies to 2,209 students. These students are assigned the standardized score on the respective test domain that they took as their overall score in language. Our main explanatory variable is the average class size in grade 3 for a given year and school. On average, class size is 20.8 for the academic years 2003 to 2006 in Saarland. Figure 4.1 illustrates the range of variation in average class size in grade 3 across as well as within schools. It is obvious that most of the variation is between schools, however, there is also a large amount of variation in average class size within schools. This is important, as we exploit only this part of variation in class size for our estimations.

Figure 4.1: Class size variation



Notes: The figure shows density plots for the total and the within-school variation in average class size in grade 3, where average class size in grade 3 is normalized to have mean zero. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

In addition to test scores, the SOE data contain a rich set of control variables. Teachers reported gender, nationality, language spoken at home, age in years, German proficiency, and learning disabilities for each student. Students also reported the number of books at home, which is a useful proxy for socio-economic family background. Ammermueller and Pischke (2009) show that the reported books at home indicator strongly correlates with a host of parental background measures such as income, education, and origin. In fact, Wößmann (2005) and Ammermueller and Pischke (2009) find it to be the single most important predictor of cognitive skills in the Third International Math and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) as well as the Programme for International Student Assessment (PISA), respectively. Unfortunately, this question was not included in the first round of testing in 2003.

The last column of Table 4.1 also reports the number of observations for each variable. For most variables the share of missing observations is less than five percent except for the books at home question. In order to preserve as much information from the data as possible we keep all observations with missing data on control variables and create an additional missing category for each variable. The lower panel of Table 4.1 illustrates the impact of the school mergers in 2005. The number of schools decreased from 258 in the year 2004 to 156 in 2005 (a change of 40%) and as a result the average number of classes increased substantially from 2.33 to 3.25 classes per school.

Table 4.2 reports descriptive statistics on the fraction of students in Saarland that were enrolled late and early in grade 1 the academic years 2001-2006. It further contains the fraction of students repeating each grade during those school years. On average, 9 percent of all students repeat a grade before fourth grade, 2.5 percent are enrolled late and 7 percent are enrolled early.

4.7 Results

4.7.1 Evidence on the Validity of the Theoretical Model

Our data allow us to test whether changes in birth cohort size lead to the predicted compositional changes in primary schools on the grade-level as discussed in section 4.3. We use administrative enrollment data for grade 1 for Saarland and regress the fraction of

Table 4.2: Descriptive statistics: Timing of school enrollment and grade repetition

	Mean (in %)
Early enrolled	7.0
Late enrolled	2.5
Grade repetition	
1st grade	3.2
2nd grade	2.9
3rd grade	2.8
4th grade	1.9

Notes: The table reports means of the listed variables for the school years 2001/2002-2006/2007. Source: Statistisches Bundesamt (2010).

students in grade 1 who were retained in grade 1 the year before, the fraction of students enrolled late, and the fraction enrolled early on the imputed cohort size for that year and school fixed effects. Panel A of Table 4.3 reports the results of these regressions. All coefficients have the expected negative sign and are statistically significant. For example, for the fraction of late enrolled students, we obtain a point estimate of -0.213. This estimate implies that if a birth cohort is increased by one student, students who have been enrolled one year too late will account for 0.213 percentage points fewer students in grade 1 in the year that this cohort is expected to enroll.

The actual instrument we use is the predicted class size based on imputed cohort size. To assess whether this instrument is also systematically related to the composition of students on the grade-level, panel B presents estimates where we use class size in grade 1 as explanatory variable and instrument it with the predicted class size based on the imputed cohort size. Again, all coefficients have the expected negative sign and are statistically significant. However, the coefficients increase substantially in size compared to panel A. For instance, an increase of one student in the predicted class size in grade 1 based on imputed cohort size is associated with a decrease in the share of students in grade 1 who

Table 4.3: Effects of cohort size on student composition

	% Late enrolled	% Early enrolled	% Repeater
	(1)	(2)	(3)
	Panel A: OLS grade composition		
Imputed cohort size	-0.213*** (0.026)	-0.164*** (0.023)	-0.045** (0.020)
	Panel B: IV grade composition		
Class size	-0.800*** (0.081)	-0.476*** (0.073)	-0.262*** (0.055)
	Panel C: OLS birth cohort composition		
Imputed cohort size	0.029 (0.025)	0.002 (0.029)	
N SchoolYearObs	871	871	871

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each cell contains results for separate, weighted regression with weights equal to total enrollment. Panel A reports estimates of the effects of imputed cohort size on the percentage of repeating, late, and early enrolled students in grade 1. Panel B reports instrumental variables estimates of average class size in grade 1 on the percentage of repeating, late, and early enrolled students in grade 1. The instrument for class size is imputed cohort size divided by the number of classes. Panel C reports estimates of the effects of imputed cohort size on the percentage of repeating, late, and early enrolled students in a birth cohort. Regressions include school and year fixed effects. Standard errors clustered at the school-level are given in parentheses. Source: Statistisches Amt des Saarlandes (2017).

were enrolled too late by 0.8 percentage points. Therefore, it appears that the compositional effects on the grade-level that arise from a cohort's size are amplified when cohort size is used in an IV framework to predict class size. It is easy to see why this is the case. Since most schools have more than one class, class size does not increase one for one with cohort size. Hence, the compositional effects in panel A are upward scaled by the inverse of the average increase in class size associated with a one-student-increase in cohort size to obtain the IV estimates.

To further check that these compositional effects result mechanically, we implement a data-generating process that is tailored to the primary school system in Saarland in terms of the size of cohorts and the fraction of retained students. Taking mean estimates from 1,000 simulations, gives similar results to those reported in panels A and B. The simulations and further discussion can be found in Appendix E and Table A4.12.

Moreover, we obtained administrative, school-level enrollment and grade retention data for all public primary schools for the 2004-2015 school years for the state of Saxony, which has retention rates in grades 1-3 that are very similar to those in Saarland. Columns 1-3 of Table A4.4 show results for Saxony analogous to those reported in Table 4.3 with similar findings. In addition, the data for Saxony contain information on the number of students who have been retained in grades 2 and 3. This allows us to explore how initial birth cohort size affects the grade-level composition of students in higher grades. In columns 4 and 5 of panel A, we, therefore, regressed the fraction of students who have been retained until grade 2 and 3 on the imputed cohort size. Columns 4 and 5 of panel B show results where the same outcomes are regressed on class size in grade 2 and 3, instrumented by the predicted class size based on the imputed cohort size. The fact that the IV estimate for class size in grade 3 in column 5 of panel B is about three times the size of the coefficient for grade 1, suggests that we can approximate the corresponding effect in grade 3 for Saarland by simply multiplying the effect in column 3, panel B of Table 4.3 by three.

The theoretical results in section 4.3 imply that instrumental variable estimates will be biased if non-retained students have skills that differ, on average, from retained, red-shirted, and early enrolled students. We next test for average skill differences between these groups. As mentioned before, our test score data only contain students' age in years. This precludes to distinguish between students who were enrolled one year late and those who were retained in primary school, as they will both appear as older than 9 years in our data. Further, we cannot distinguish between students who were enrolled one year early and those who were born between May and June but enrolled on time. Instead, we use data from the NEPS starting cohort 2, which is a representative sample of primary school children from Germany. The NEPS contains several skill measures, information on whether a child has been retained, and the timing of school enrollment.[99] Thus, it allows

[99] More information on this dataset and how we constructed the skill measures is provided in Appendix

identifying each group of students. Table 4.4 reports results from regressions of measures of language, math and cognitive skills on dummy variables for each separate group of students. As expected, retained and late enrolling children score lower on all three skill tests. The point estimate for grade repeaters for math implies that students who have been retained in the past have 0.9 SD lower math skills than regular students. Surprisingly, students who were enrolled early do not differ significantly from regular students in terms of their skills. Therefore, we expect the potential bias introduced by early enrollment to be of little concern.[100]

With the results from Tables 4.3 and 4.4 we can perform a simple exercise to quantify the expected bias resulting from grade retention in class size estimates based on the IV approach. In Equation (4.15) we see that the bias is additive and equals the product of $(\theta - \delta)$ and ρ_{IV}.[101] Consequently, we simply multiply the expected compositional effect of birth cohort size on the fraction of students of a particular group in grade 3 (ρ_{IV}) with the average test score difference between that group and the group of students who reach grade 3 on time $(\theta - \delta)$. Under the assumption that the compositional effect in grade 1 can be linearly extrapolated to grade 3, this yields values of 0.564 ($= 3 \times 0.262 \times 0.717$) SD and 0.715 ($= 3 \times 0.262 \times 0.910$) SD for retained students for language and math, respectively.[102] For the full bias, we add the bias arising from late enrolled students: 0.175 ($= 0.8 \times 0.219$) SD for language and .227 ($= 0.8 \times 0.284$) SD for math. Combining these results, we expect the bias from compositional effects to decrease estimates of a 10-student-reduction in class size between grades 1-3 on test scores in grade 3 by 0.074 SD for language and 0.094 SD for math.

B.

[100] Another potential concern are students who skip a grade. Table 4.4 shows that these students have up to 0.96 SD better skills than regular students. However, the share of students who skip a grade before grade 3 is very low. There are no official data on grade skipping for Saarland, but NEPS data show that less than 0.6 percent of students skip a grade before grade 3 in Germany.

[101] We do not take into account the bias resulting from attenuation here. Hence, we get a lower bound of the true size of the bias.

[102] The value 0.262 comes from column 3 of panel B in Table 4.3. The second value, 0.717, is from row 3 and column 1 in Table 4.4. The second value for math comes from the second row of column 2 in Table 4.4. Our results for Saxony, where (similar to Saarland) the grade retentions rates are almost constant in grades 1-3, indicate that the compositional effect in grade 3 can be approximated by multiplying the effect in grade 1 by 3.

Table 4.4: Differences in skills of late-, early enrolled, and grade repeating students

	Language	Math	Cognition
	(1)	(2)	(3)
Late enrolled	-0.219***	-0.284***	-0.160***
	(0.048)	(0.044)	(0.050)
Grade repeater	-0.717***	-0.910***	-0.525***
	(0.059)	(0.056)	(0.079)
Early enrolled	-0.031	0.047	0.022
	(0.046)	(0.048)	(0.045)
Grade skipper	0.940***	0.963***	0.507***
	(0.165)	(0.115)	(0.115)
N	5727	6373	5153

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each column contains the coefficients for a regression of the respective skill on the variables listed in the rows. Robust standard errors are given in parentheses. Source: NEPS Data, Data Version SC2: 6.0.1.

4.7.2 Validity of Birth Cohort Size Variation

The key assumption of our estimation approach described in section 4.5 is that within schools, changes in birth cohort size are unrelated with ability levels of cohorts and the thresholds that determine grade retention, redshirting, and early school enrollment. We use two approaches to check the validity of this assumption. First, we test whether birth cohort size is related to the fraction of students from a birth cohort who are enrolled late or early, by regressing these fractions on cohort size and school fixed effects. Panel C of Table 4.3 reports the results of these regressions. Reassuringly, the results indicate that early and late enrollment is balanced with respect to birth cohort size.[103] This lends support to the hypothesis that birth cohort size is not related to student ability or the thresholds that determine early- or late school enrollment. In light of our discussion of the results in panel A, any correlation between initial cohort size and the composition of students in higher grades seems to be driven by mechanical relationships rather than

[103] We omit the result for the fraction of students who repeat a grade in column 3. The reason is that if class size has a negative impact on student achievement, we expect a significant positive effect of cohort size on retention rates even if cohort size is unrelated to the composition of cohorts. This will be discussed further below.

Table 4.5: Balancing tests

	Explanatory variables				
	Test Score Equations		Balancing Test		
	Language	Math	Imputed Cohort Size		
Dependent variables	(1)	(2)	(3)	(4)	(5)
Insufficient German Proficiency	-0.0732***	-0.0511***	0.0001	-0.0008**	-0.0004
	(0.0028)	(0.0026)	(0.0001)	(0.0003)	(0.0003)
Older than 9 at test date	-0.0877***	-0.0688***	0.0001	-0.0009***	-0.0004
	(0.0026)	(0.0025)	(0.0002)	(0.0003)	(0.0003)
Younger than 9 at test date	0.0308***	0.0215***	-0.0002*	-0.0010***	-0.0009**
	(0.0019)	(0.0020)	(0.0001)	(0.0004)	(0.0004)
Age in years	-0.1340***	-0.1013***	0.0003	0.0001	0.0004
	(0.0042)	(0.0040)	(0.0003)	(0.0006)	(0.0006)
Male	-0.0521***	0.0369***	-0.0002	0.0007*	0.0008*
	(0.0029)	(0.0028)	(0.0001)	(0.0004)	(0.0005)
Migration Background	-0.0827***	-0.0564***	0.0012***	-0.0004	-0.0001
	(0.0052)	(0.0041)	(0.0004)	(0.0004)	(0.0004)
Non-native German Speaker	-0.0851***	-0.0581***	0.0011***	-0.0006	-0.0003
	(0.0054)	(0.0043)	(0.0004)	(0.0005)	(0.0005)
Reported books at home					
Index	0.3129***	0.2569***	-0.0024**	-0.0001	-0.0004
	(0.0104)	(0.0103)	(0.0011)	(0.0018)	(0.0015)
None or few books	-0.0474***	-0.0372***	0.0003	-0.0006**	-0.0003
	(0.0030)	(0.0026)	(0.0002)	(0.0003)	(0.0002)
Enough to fill one shelf	-0.0515***	-0.0438***	0.0005***	0.0007	0.0006
	(0.0024)	(0.0022)	(0.0002)	(0.0005)	(0.0005)
Enough to fill one bookcase	0.0341***	0.0243***	-0.0001	0.0000	0.0001
	(0.0028)	(0.0028)	(0.0002)	(0.0005)	(0.0005)
Enough to fill two bookcases	0.0662***	0.0572***	-0.0006**	-0.0003	-0.0003
	(0.0034)	(0.0036)	(0.0003)	(0.0006)	(0.0006)
Dyscalculia	-0.0401***	-0.0461***	0.0001	-0.0007	-0.0000
	(0.0024)	(0.0027)	(0.0001)	(0.0006)	(0.0006)
Dyslexia	-0.0781***	-0.0467***	-0.0001	0.0002	0.0005*
	(0.0032)	(0.0024)	(0.0001)	(0.0003)	(0.0003)
Rural community	0.1097***	0.1026***	-0.0108***		
	(0.0198)	(0.0191)	(0.0032)		
Problematic school district	-0.0771***	-0.0675***	0.0046***		
	(0.0109)	(0.0100)	(0.0015)		
N Cluster	156	156	156	156	156
Year FE	Yes	Yes	Yes	Yes	Yes
School FE				Yes	Yes
Cohort adjusted					Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each cell contains results for a separate regression. Columns 1-3 report results of OLS regressions of the variables listed in the rows on the listed characteristics in the column header. All regressions include cohort fixed effects. Column 4 reports results of OLS regressions of the same variables but also controlling for school fixed effects. Column 5 reports results where students who are older than 9 years are assigned to the cohort of the previous year. Robust standard errors clustered at the school-level are given in parentheses. Index refers to a linear index of the reported books at home. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

correlations between the size and initial composition of birth cohorts.

In a second approach, we check whether student characteristics are balanced with respect to birth cohort size drawing on the student-level data. In Table 4.5, each cell contains the result from a separate regression of the student characteristic listed in the row on the variable listed in the column. The first two columns show that all variables we consider are highly relevant predictors of language and math test scores and have the expected signs. Columns 3-5 report the results of regressing student characteristics on imputed cohort size. Almost half of the coefficients in column 3 are significant, which is evidence for considerable across-school-sorting of students with respect to cohort size. Once we condition on school fixed effects in column 4, most coefficients turn insignificant. However, consistent with our model's prediction of a negative relationship between initial cohort size and the share of students held back or enrolled early on the grade-level, the coefficients for being older and younger than typical third graders are significant and negative.[104] More generally, any significant effects in column 4 could be the result of compositional changes caused by initial cohort size. This can explain the significant negative coefficients for limited Germany proficiency and reporting none or few books at home as these are characteristics that correlate strongly with having been enrolled late or retained.

To actually test whether the initial birth cohort composition is balanced with respect to cohort size, we need to assign students to their respective birth cohorts. To this end, we reassign students who report being older than 9 years to the cohort of the previous year. The results of these regressions are reported in column 5.[105] In contrast to column 4, the significant associations of cohort size with limited German proficiency, being older than 9 years, and reporting none or few books at home disappear. These results indicate that within schools student characteristics of birth cohorts are balanced with respect to birth cohort size.[106]

[104] We suspect that these patterns were not discovered in previous within-school studies which performed similar balancing tests such as Wößmann and West (2006) because they only checked for a linear relationship between age and class size. Note that in column 4 there is no significant effect for cohort size on age in years despite the significant negative effects for being older and younger than 9.

[105] Since we lack data for 2002, we cannot assign grade repeaters and late enrolled students to the birth cohort that reaches 3rd grade regularly in 2003. Hence, we drop this cohort for the regressions in column 5. However, the results are very similar when this cohort is included. Further, we refrain from assigning students who report being younger than 9 to next year's birth cohort because most of these students were born between May and June and, hence, reached grade 3 on schedule rather than being enrolled early. This explains why we still find significant effects for being younger than 9 in column 5.

[106] As expected when running a number of regressions testing multiple hypotheses, some coefficients are weakly statistically significant. In the absence of any correlation between birth cohort size and student characteristics we would expect 10 percent of coefficients to be statistically significant at the 10 percent significance level. The share of significant coefficients (not counting the coefficient for being younger than 9) in column 5 is, at 14 percent, only slightly above this expected value.

Table 4.6: The effects of insufficient German proficiency on number of classes and class size

	# classes	Class size
	(1)	(2)
Insufficient German proficiency	0.017**	-0.169**
	(0.007)	(0.074)
Enrollment grade 1	0.040***	0.035**
	(0.002)	(0.016)
School FE	Yes	Yes
N Students	38415	38415

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each column contains results for a separate regressions. Standard errors clustered at the combined school-level are given in parentheses. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

We next examine whether the lower class size thresholds for grades with more students with insufficient German proficiency could lead to a positive bias in within-school estimates of class size effects. Table 4.6, column 1 reports results where we regress the number of classes in grade 3 on an indicator for insufficient German proficiency measured in grade 3, total enrollment in grade 1, and school fixed effects. The positive coefficient for German proficiency indicates that grades with more students not proficient in German have significantly more classes holding enrollment constant. This, in turn, implies that class size for these students is about 0.169 students smaller than it is for students proficient in German from the same school with the same number of students in a grade; see column 2. Because of this feature of the data, we will control for German proficiency in some of the analyses below.

4.7.3 Class Size Effects

In this section, we turn to reporting our class size effects. Table 4.7 reports first stage coefficients for our instrument, predicted class size based on imputed cohort size, on average class size in grade 3. As expected, the instrument is a strong predictor of class size and the F-statistic is above 170 for all specifications. Our results indicate that a one-

Table 4.7: First stage estimates

	Class size in grade 3			
	(1)	(2)	(3)	(4)
Class size predicted by imputed cohort size	0.446*** (0.034)	0.446*** (0.034)	0.446*** (0.034)	0.446*** (0.034)
School FE	Yes	Yes	Yes	Yes
Age Controls		Yes	Yes	Yes
Insufficient German Proficiency			Yes	Yes
Individual Controls				Yes
N	38415	38415	38415	38415
R^2	0.345	0.345	0.346	0.347
F-Test	172	172	172	174

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. The table shows estimates of the effects of class size predicted by imputed cohort size on class size in grade 3. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

student-increase in predicted class size based on imputed cohort size leads approximately to a 0.45-student-increase in class size in grade 3.

Tables 4.8 contains our main results for the empirical model in Equation (4.18). We run separate regressions for language and math to be able to draw subject-specific conclusions. Column 5 reports results from IV regressions where we only control for school and year fixed effects.[107] The point estimates in both subjects are negative but not statistically significant. Our discussion of Equation (4.14) suggests, however, that these estimates might suffer from a positive bias because of the correlation between initial cohort size and the composition of students in higher grades. Once we include age controls in column 6, the IV estimates for language and math almost double in absolute size. This is consistent with the comparison of equations (4.14) and (4.15). The implied upward bias in class size estimates without age controls is 0.071 SD for language and 0.06 SD for math, which is in the ballpark of the predicted bias based on our theoretical model.[108]

[107] The full regression results are reported in Tables A4.5 -A4.6 in Appendix A.

[108] In section 4.7.1, we calculated a bias for a one-student-increase in class size of 0.074 for language and 0.094 SD for math. As discussed in section 4.5, however, the differences in coefficients in columns 5 and 6 are likely to understate any bias resulting from holding back poorly performing students. This is because we only condition on a proxy for whether or not a student has been held back in the past, which does not

Table 4.8: Main results: The effect of class size on test scores

	OLS				IV			
	Avg. class size grade 3				IV: Imputed cohort size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Language	-0.0159***	-0.0178***	-0.0202***	-0.0199***	-0.0074	-0.0145*	-0.0189**	-0.0191**
$[N = 37,847]$	(0.0045)	(0.0044)	(0.0052)	(0.0050)	(0.0085)	(0.0085)	(0.0095)	(0.0092)
Math	-0.0112	-0.0127*	-0.0143**	-0.0140**	-0.0061	-0.0121	-0.0150	-0.0140
$[N = 36,845]$	(0.0068)	(0.0068)	(0.0072)	(0.0070)	(0.0108)	(0.0108)	(0.0111)	(0.0110)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls		Yes	Yes	Yes		Yes	Yes	Yes
Insufficient German proficiency			Yes	Yes			Yes	Yes
Individual controls				Yes				Yes
N Cluster	156	156	156	156	156	156	156	156
N SchoolYearObs	828	828	828	828	828	828	156	156

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each cell contains results for a separate regression. Columns 1-4 report OLS estimates of class size in grade 3 on language and math. Columns 5-8 report estimates of class size in grade 3 where class size is instrumented by predicted class size based on imputed cohort size. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

The differences between estimates in columns 5 and 6 are not statistically significant and only the language effect turns weakly significant when we control for age. Nevertheless, these findings are suggestive of a potentially substantial bias in IV estimates of class size effects in school systems where students can be retained or redshirted.

Because students with insufficient German proficiency are, on average, placed in smaller classes in Saarland (see the discussion in section 4.5 and Table 4.6), the results in column 6 are likely still upward biased. Controlling for German proficiency in column 7 confirms this. Class size coefficients for both subjects become considerably more negative and the language effect turns significant at the five percent level. Including further controls such as a gender dummy or the reported number of books at home in column 8, however, makes little difference to the results. This suggests that any bias in our within-school estimates seems to be driven either by compositional effects arising from held back students or the lower class size threshold for students with insufficient German proficiency. Once we control for these confounding effects, the class size coefficient for language implies a statistically significant test score increase of 0.0191 SD for a one-student-decrease in class size from grade 1 until grade 3. For math, the corresponding effect size is 0.014 SD, although the estimate is not statistically significant.

fully eliminate the bias resulting from these students. Therefore the implied size of the bias in Table 4.8 is a lower bound, explaining why it is slightly smaller than what we predicted.

The OLS results in columns 1-4 follow the same pattern as the IV results. Estimated class size effects become more negative as we control for age and insufficient German proficiency, but do not change with the inclusion of further controls. However, estimates for language and math in column 1 without any age controls are substantially larger in absolute size than the corresponding IV estimates. For language, the effect is significant at the one percent level. The inclusion of age controls only modestly increases class size estimates in size in column 2. This could point to a lower compositional bias in within-school designs that regress test scores directly on class size compared to the IV approach. One possible explanation is that held back students increase the size of the class they join after having been held back. A positive correlation between class size and the share of retained students ensues, which offsets part of the negative correlation between class size and the share of held back students discussed before.[109] Notably, with controls for age and German proficiency the OLS results in column 4 are very similar to the IV results in column 8. Durbin-Wu-Hausman tests fail to reject the null of no endogeneity in all IV specifications in columns 5-8 for language and math. Therefore, the overall conclusion is that the OLS results seem to be robust to the potential bias ι_{OLS} in Equation (4.17) in our setting. The substantially smaller OLS standard errors render estimates of class size effects for language and math in columns 3-4 statistically significant at the at the one and five percent level, respectively. We view this as strong evidence for a negative impact of class size on students' test scores.

Importantly, the true magnitude of the class size effects is likely to be larger than the estimates presented here. Imperfect proxies for retention status and German proficiency leave some room for upward bias in our estimates. Further, equations (4.15) and (4.17) imply that the estimates in Table 4.8 are attenuated because class size in grade 3 is not perfectly correlated with the class size students experienced in grades 1 and 2.[110]

[109] Unfortunately, comparing ρ_{IV} and ρ_{OLS} in equations (4.14) and (4.16) does not allow us to conclude whether the composition bias should be larger for IV or OLS. This is because ρ_{OLS} is a function of the second moments of the shocks to ability levels and grade retention thresholds (see Equation (D4.19) in Appendix D), which cannot be identified.

[110] Table A4.7 reports estimates for different specifications using either average class size in grade 1, grade 2, or the average of grades 1-3 as explanatory variables. OLS and IV results for both subjects exhibit a monotonic pattern. Estimated class size effects appear to decrease in absolute size if test scores are regressed on class size from lower grades and results for the average class size in grades 1-3 fall somewhere

As a robustness check we also estimate models in which we include separate fixed effects for each school and number of classes combination instead of school fixed effects. This amounts to identifying the class size effect only by within-school-variation in class size that is caused by changes in cohort size while holding the number of classes constant. These specifications more closely follow Hoxby (2000) who conditions on the expected number of classes and should be less prone to bias caused by the addition of newly hired teachers whenever a school changes the number of classes as discussed in Gilraine (2018). Columns 3 and 6 of Table A4.9 report the results of these regressions. Although we lose considerable variation in class size that is driven by schools adding or removing a class, the estimates are qualitatively very similar to the results in Table 4.8. However, while the OLS estimates are still significant, the IV results lose statistical significance because of a substantial increase in standard errors.

Our balancing tests in Table 4.5 indicate that the within-school variation in cohort size we use to identify class size effects is unrelated to observed determinants of student achievement in our data. Nevertheless, one may still be concerned that our estimates are picking up school-specific trends in cohort size. If, for example, there is an inflow of young families moving into a school's catchment area, this might bias the result if children from these families differ on average from other children in the catchment area. Although we expect that our balancing results should indicate compositional changes in the student population that correlate with cohort size, we further check that school-specific trends in unobserved determinants of student achievement do not drive our class size effects. The drawback is that the within-school variation of class size is substantially reduced if we take out linear trends in a panel with only four years.[111] In fact, any school with less than

between the results for grade 1 and grade 3. This is consistent with the notion that for students who enter a class after grade 1 (e.g. because they have been retained or switched schools), the class size for grade 1 of the class in which we observe them in grade 3 is an erroneous measure of their previous class size. Note that we do not observe when a students has been held back or switched school. Therefore, we cannot assign these students to their previous classes. The fact that test scores are measured at the end of grade 3 and retention and most school switches happen at the end of the school year ensures that, except for some rare cases, all students should have experienced at least the class size we observe in grade 3. Hence, we expect measurement error to be minimized by using class size in grade 3 as the explanatory variable.

[111] Hoxby (2000) estimates more flexible time trends with a quartic in time. However, our data have only panels with at most four years. For this short of a period, any trend should be adequately summarized by a linear trend.

three years of data has to be dropped from the analysis. Hence we lose about 60 percent of all observations.[112] The results of these regressions are reported in columns 2 and 5 of Table A4.9. The loss of observations and variation in class size roughly doubles the standard errors in these regressions. Hence, most coefficients turn insignificant. However, all coefficients increase in absolute size, which indicates that, if anything, school-specific trends in cohort size seem to be positively correlated with student achievement. This is in line with an explanation based on the inflow of young families with higher socio-economic status into a school's catchment area causing an increase in cohort size. As this would bias our class size effects positively, we expect our estimates without school-specific linear trends in Table 4.8 to provide lower bounds on the true class size effect.

4.7.3.1 Non-Linear Effects

So far, we have assumed linear class size effects, i.e. that a one-student-increase in class size has the same effect in smaller and larger classes. This may not be a sensible assumption. We may think of a situation in which class size effects increase in larger classes; for instance if the growing potential for disturbances in larger classes is partly offset by more efficient instruction up until a certain threshold, because a "critical mass" of good students is required for fruitful discussions. The same may happen if the potential for classroom disturbances grows exponentially in larger classes, for example because a "critical mass" of problematic students is reached and their disturbances reinforce each other. Alternatively, we could think of a situation in which the potential for disturbances becomes flatter as classes grow larger, because the addition of more problematic students makes a smaller difference percentage-wise in larger classes. This line of argument is used by Hoxby (2000) to motivate a level-log model specification. While this is by no means an exhaustive list of potential explanations for non-linear class size effects, it serves to illustrate that a variety of (potentially countervailing) forces may be at work in classrooms that make studying non-linearities worthwhile.

[112] Recall that two-thirds of schools were merged prior to the 2005 school year resulting in only two years of data for schools that were eventually merged before the consolidation and two years of data for the combined schools after the consolidation.

Table 4.9: Spline regressions

	17.5	18.5	19.5	20.5	21.5	22.5	23.5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Panel A: Language			
Class size $<$ knot	0.0174	0.0152	0.0140	0.0056	-0.0041	-0.0109	-0.0158**
	(0.0217)	(0.0161)	(0.0128)	(0.0105)	(0.0085)	(0.0073)	(0.0068)
Class size $\geq$ knot	-0.0310***	-0.0351***	-0.0420***	-0.0483***	-0.0531***	-0.0586***	-0.0638***
	(0.0061)	(0.0067)	(0.0076)	(0.0091)	(0.0107)	(0.0130)	(0.0171)
N	37847	37847	37847	37847	37847	37847	37847
R^2	0.263	0.263	0.264	0.264	0.263	0.263	0.263
				Panel B: Math			
Class size $<$ knot	0.0058	0.0093	0.0110	0.0064	0.0027	-0.0039	-0.0095
	(0.0226)	(0.0168)	(0.0139)	(0.0123)	(0.0107)	(0.0093)	(0.0085)
Class size $\geq$ knot	-0.0226***	-0.0261***	-0.0321***	-0.0384***	-0.0482***	-0.0551***	-0.0594**
	(0.0086)	(0.0092)	(0.0104)	(0.0126)	(0.0156)	(0.0201)	(0.0273)
N	36845	36845	36845	36845	36845	36845	36845
R^2	0.157	0.157	0.158	0.158	0.158	0.158	0.158
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. The table reports OLS results for different linear spline specifications with a single knot the position of which is indicated in the column header. The coefficients measure class size effects for the specified interval in the first column. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include dummies for age in years, gender, number of books at home, migration background, native language, and an indicator of insufficient German proficiency. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

In Table 4.9 we report estimates from several spline regressions with a single knot placed at different class size values, thereby allowing class size effects to differ between small and large classes. Since our results above indicate that OLS and IV specifications yield similar results once we condition on age and German proficiency, we only report the more efficient OLS results.[113] Throughout all specifications, there is clear evidence for non-linear effects. Specifically, large negative class size effects are predominantly evident in larger classes. For instance, the estimated effect for classes larger than 20.5 students indicates a reduction in language test scores of 0.0483 SD for each additional student, while the effect for classes smaller than 20.5 is statistically insignificant. Panel B shows the same pattern of basically zero effects in small classes and large negative effects in larger classes for mathematics.[114]

[113] The IV results are reported Table A4.8. They are very similar to the OLS results, albeit noisier.

[114] As before, we also carry out robustness checks, such as including school-number of classes combination fixed effects, and school specific linear trends. Table A4.10 in Appendix A reports results for the spline

The finding of non-linear effects might have important implications for the empirical class size literature, which generally uses class size measures aggregated at the grade-level or even school district level. Since class size effects operate at the individual class level, using more aggregate measures of class size could not only result in larger standard errors, but also inconsistent estimates when these effects are non-linear. Hence, we speculate that using class size variation at the grade-level might underestimate the class size effect if the effect is actually non-linear and class size is very heterogeneous within grades. This result may help reconcile some of the zero findings in the literature by studies that measure class size at the grade-level (e.g. Angrist et al., 2017b,a; Wößmann and West, 2006) and even more so for the study by Hoxby (2000) which uses variation in class size at the school-district-level. The level of aggregation as one possible explanation for different findings across studies is also consistent with those studies that measure the effect of class size at the class level by Krueger (1999), Urquiola (2006) and Bressoux et al. (2009): these studies find large and significant class size effects.[115]

4.7.3.2 Effect Heterogeneity

In our specifications in Tables 4.8 and 4.9 we implicitly assume that all students are similarly affected by class size. Krueger (1999), however, finds more pronounced effects of class size reductions for disadvantaged groups. We test for these sources of heterogeneity by interacting the class size variable with a set of indicator variables for gender, being too old for grade 3, reporting few books at home, migration background, insufficient German proficiency, reading disorder (dyslexia), and learning disability in math (dyscalculia). Table 4.10 shows the coefficients of these seven interactions.[116] In line with the hypothesis that disadvantaged students are harmed most by larger classes, all interaction terms pertaining to disadvantaged groups of students, are negative and most are statistically

specification with a knot placed at 20.5. The results are qualitatively very similar, but as before, standard errors increase substantially.

[115] The results in Leuven et al. (2008) provide some evidence against this hypothesis as they find no significant class size effects for Norwegian schools with only one class per grade where average class size equals actual class size. However, their study investigates the effects of class size in lower secondary school and class size effects are generally thought to be larger in primary school.

[116] Since the IV results are very similar we only report OLS results. For the IV results, see Table A4.11 in Appendix A.

significant at the one percent level. Additional evidence comes from the pattern of the interaction terms for dyslexia and dyscalculia. If students react more strongly to class size in subjects where they are at a disadvantage, we should expect larger effects for dyslexic students in language compared to math and vice versa for students with dyscalculia. This is exactly what we find in columns 6 and 7 in panels A and B. Moreover, the interaction term for dyslexia is larger than the one for dyscalculia in language and vice versa in math, which we would also expect.

More importantly, the estimated class size effects for disadvantaged students are very large in magnitude: for example, the coefficient for insufficient German proficiency suggests that one more student in class decreases language and math test scores of students not proficient in German by 0.053 and 0.037 SD, respectively. Overall, these results reveal that our specifications in Tables 4.8 and 4.9 mask some marked effect heterogeneity for certain groups of students. Compared to non-disadvantaged students, class size effects seem to be two to four times larger for students who can be expected to be at a disadvantage either because of their migration status, insufficient German proficiency, learning disabilities, or lower academic skills as evident from having been held back a grade.

4.7.3.3 Effects on Grade Retention

If class size has a negative effect on student achievement, it can also be expected to increase the probability of being retained. To explore this, we use administrative school-level data on the number of grade repeaters in grade 1 for the 2001-2004 academic years.[117] We follow the same methodological approach as above, but now regress the share of students who repeat grade 1 in year t on class size in grade 1 in year $t-1$ and school fixed effects. Since we do not have grade repetition information at the student level, we conduct the analysis at the school-year level. Column 1 in Table 4.11 reports the OLS estimate of this regression and column 3 reports the IV estimate, where average class size in grade 1 is instrumented with predicted class size based on imputed cohort

[117] Note that we have to discard data for the year 2004 for all schools that were merged in 2005. The reason for this is that we do not observe the number of students who entered first grade in 2004 and repeated the same grade in 2005 since we only have that information on the consolidated school-level for 2005. We also have to discard data for the year 2000 because we cannot impute cohort size for that year as we do not observe the number of students who were enrolled too early in 1999.

Table 4.10: Heterogeneity OLS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Panel A: Language			
Avg. class size grade 3	-0.021***	-0.018***	-0.019***	-0.018***	-0.018***	-0.017***	-0.019***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
× female	0.003						
	(0.003)						
× older than 9 years		-0.016***					
		(0.006)					
× few books			-0.007				
			(0.004)				
× migration background				-0.014***			
				(0.005)			
× insufficient German proficiency					-0.035***		
					(0.001)		
× dyslexia						-0.041***	
						(0.001)	
× dyscalculia							-0.032***
							(0.001)
N	36845	36845	36845	36845	36845	36845	36845
				Panel B: Math			
Avg. class size grade 3	-0.013*	-0.012*	-0.013*	-0.012*	-0.013*	-0.013*	-0.013*
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
× female	-0.002						
	(0.004)						
× older than 9 years		-0.015***					
		(0.005)					
× few books			-0.005				
			(0.005)				
× migration background				-0.013**			
				(0.005)			
× insufficient German proficiency					-0.024***		
					(0.001)		
× dyslexia						-0.023***	
						(0.001)	
× dyscalculia							-0.044***
							(0.001)
N	37847	37847	37847	37847	37847	37847	37847
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Limited German proficiency	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports OLS results where each column panels A and B contains the results for a separate regression with the same specification as that of column 3 in Table 4.8, except that the class size variable is interacted with an indicator variable for the individual student characteristics. Few books is a dummy for reporting enough books to fill one shelf or less. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include age in years, gender, number of books at home, migration background, learning disabilities, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

size. Both estimates indicate that larger classes in grade 1 increase the share of students who are retained in first grade significantly.

Given the discussion in section 4.3, however, the estimate in column 3 may be biased because predicted class size based on imputed cohort size is mechanically related to the composition of students in grade 1. Here, the bias should go in the opposite direction as above, i.e. we should overestimate the positive effect of class size on grade retention rates. To see this, note that large cohorts should have a smaller share of students in grade 1 who have been retained in the past. Since students in Saarland are rarely retained more than once in primary school, students who have not been retained before are more likely to be retained.[118] Since these students account for a larger share in larger cohorts within a school, this should lead to a positive association between cohort size (and hence class size) and the share of retained students even in the absence of any "pure class size effect." To alleviate this source of bias, we also estimate regressions where we use the share of retained students only among the students who have not been retained before as outcome variable, instead of the fraction of retained students in grade 1. The results of these regressions are reported in columns 2 and 4. As expected, the IV estimate decreases slightly but not substantially.[119] A one-student-increase in class size is associated with an increase in the fraction of repeaters in grade 1 of around 0.152 percentage points. Given that only 2.3 percent of all students repeat grade 1, this is an increase of almost 7 percent.[120] Against the background of the rather small intervention of a one-student-change, this is a very large effect. These estimates confirm earlier results by Argaw and Puhani (2018) both in substance and in size in a longer panel (four cohorts versus two) and in a different German state (Saarland versus Hesse).

Importantly, this finding may have also implications for RDDs based on maximum class size rules. As retention rates increase with class size, marginal students with low

[118] Students are rarely retained more than once in primary school because if they are, they are classified as students with special needs and then are transferred to special schools.

[119] The OLS estimate increases marginally. This is also to be expected since an increase in class size caused by an inflow of retained students from the previous year also decreases the share of students who have not been retained in the past (hence who are more likely to be retained). The OLS estimate may pick up this negative spurious effect of class size on the retention rate. Using the share of retained students among students who have not been retained before as the outcome, however, should alleviate this source of bias and, therefore, increase the OLS estimate.

[120] The retention rate of 2.3 percent is the average retention rate in grade 1 for the estimation sample. Hence, it differs slightly from the value reported in Table 4.1, which is the the population average for the 2001-2006 academic years.

Table 4.11: The effect of class size on grade repetition

	OLS		IV	
	(1)	(2)	(3)	(4)
Repeater in %	0.106**	0.110**	0.157***	0.152***
	(0.044)	(0.045)	(0.053)	(0.053)
% - change	4.80	4.95	7.09	6.87
Year FE	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes
Adjusted Repeater	No	Yes	No	Yes
N School-years	872	872	871	871
F-Test			1135	1135

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. The table reports estimates of the effect of class size in 1st grade on grade repetition rates in 1st grade. The outcome variable in columns 2 and 4 is the grade repetition rate for students who have not been retained before. The instrument in Columns 3 to 4 is the predicted class size based on imputed cohort size. The unit of observation is the school-cohort-level. Regressions are weighted by total enrollment. The sample includes all schools with at least one class per grade for the academic years 2001/2002 - 2004/2005. F-Test reports the F-test for the excluded instrument. Standard errors clustered at the school-level are given in parentheses. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

academic skills should have a higher likelihood of being retained in large classes just below the class-size threshold as compared to if they were in smaller classes just above. Class size estimates based on a comparison of student test scores between these classes in higher grades could therefore suffer from a form of survivorship bias. A back-of-the-envelope calculation for schools with a class size cap of 29 and enrollment between 29-30 students yields that an RDD estimate for the effect of a 10-student increase in class size would be upward biased by 3.3 and 4.2 percent of a SD for language and math, respectively.[121]

[121] To get those values, note that class size in schools with 29 students is 29 and 15 in schools with 30 students. If we abstract from the composition effects discussed in section 4.3 and assume that the class size effect on grade retention of 0.152 for grade 1 (from Table 4.11) can be linearly extrapolated to grade 3, we get a difference in retention rates by grade 3 between classes that were initially of size 29 and 15 equal to 6.384 percentage points ($= 14 \times 0.152 \times 3$). Multiplying this by the average difference in test scores between non-retained and retained students in Table 4.4 and dividing by the class size difference, yields an RDD estimate of 0.0033 SD ($= 3 \times 0.00152 \times 0.717$) and 0.0042 SD ($= 3 \times 0.00152 \times 0.91$) for language and math respectively. However, as most RDD designs have to use wider bandwidths, schools with sizable enrollment differences are compared. This could make these estimates also susceptible to the

4.8 Conclusion

Class size is a central lever for educational policy-makers as teachers' salaries make up the largest share of education spending. However, the literature remains largely inconclusive as to whether smaller classes are beneficial for student achievement. While the results from the famous randomized experiment in Tennessee (STAR) suggest that smaller classes are beneficial in terms of test scores (Krueger and Whitmore, 2001), studies using quasi-experimental approaches to identify causal effects differ substantially in their conclusions.

The theoretical model developed in this paper points out a positive bias inherent in class size estimates from standard within-school designs in school systems that allow for redshirting or grade retention. We provide important insights into the cause, consequences and solutions of this bias, which has, to the best of our knowledge, been ignored to date. Our model predicts that even if within-school changes in birth cohort size are unrelated to the initial composition of cohorts, this is not the case for the actual grade-level composition. The reason is that retaining poorly performing students mechanically causes larger birth cohorts to be in grades with a smaller share of students who have been retained before. The resulting bias may help reconcile the empirical puzzle that studies relying on idiosyncratic variation in cohort size in school systems that allow for grade retention and redshirting (e.g. Hoxby, 2000; Cho et al., 2012) mostly find no or considerably smaller effects than the experimental studies based on Project STAR. Furthermore, we provide a simple solution to this problem — controlling for whether or not a student has been held back a grade in the past — that produces a lower bound on the class size effect.

In the empirical part of this paper, we show that the two main predictions of our theoretical model find support in data on German primary schools. First, while balancing tests show the characteristics of students from the same birth cohort to be unrelated to the size of a birth cohort, we do find significant associations between birth cohort size and student characteristics at the grade-level. Second, when we estimate class size ef-

type of composition bias laid out in section 4.3. An analysis of how this affects RDD estimates is beyond the scope of this paper, but something we plan to investigate in future research.

fects with a within-school design and instrument class size in grade 3 by predicted class size based on imputed cohort size, we find that introducing a proxy for whether or not a student has been retained or redshirted leads to the expected movement in coefficients. On average, we find that a one-student-decrease in class size in grades 1-3 improves language and math test scores at the end of grade 3 by around 1.9 and 1.4 percent of a standard deviation, respectively. However, these average effects mask a significant degree of heterogeneity. Disadvantaged students seem to benefit two to four times as much from smaller classes than these average effects would suggest. Further, class size effects appear to be non-linear, with larger effects in large classes and no effects in small ones.

Our results have important policy implications. First, increasing class size to reduce public spending comes at the cost of lower student achievement. These costs are particularly large in larger classes. However, since we find little evidence of class size effects in smaller classes, the results suggest that class size may be increased up to a certain size without adversely affecting achievement. Second, larger benefits of smaller classes for disadvantaged children warrant the use of progressive maximum class size rules.

Appendix A: Figures and Tables

Table A4.1: Summary of within-school and between-cohort studies

						School system allows	
Study	Country	Grade at test	Outcome	Significant effect	Level of data aggregation	Grade retention	Late school enrollment
Hoxby (2000)	US	4/6	test scores	no	school-district	yes	yes
Rivkin et al (2005)	US	3-7	test scores	yes	student	yes	yes
Wößmann (2005)	EUR*	7-8	test scores	mostly no	student	mostly yes	mostly yes
Jakubowski & Sakowski (2006)	POL	6	test scores	yes	class	yes	yes
Wößmann & West (2006)	EUR†	7-8	test scores	mostly no	student	mostly yes	mostly yes
Leuven et al (2008)	NOR	7-9	test scores	no	student	no	yes
Jepsen & Rivkin (2009)	US	2-4	test scores	yes	school	yes	yes
Heinesen (2010)	DNK	10	GPA	yes	student	yes	yes
Cho et al (2012)	US	3/5	test scores	yes	school-district	yes	Yes
Gary-Bobo & Mahjoub (2013)	FRA	6-9	grade retention	yes	student	yes	yes
Denny & Oppedisano (2013)	US/UK	9-11	test scores	yes (opposite sign)	student	yes/no	yes/no

Notes: US=United States; EUR=European countries; POL=Poland; NOR=Norway; DNK=Denmark; FRA=France; UK=United Kingdom; *=15 European countries;†=10 European countries + Singapore. Significant effect refers to negative class size coefficients that are significant at the 5 percent level. Level of data aggregation refers to the level at which the outcome variables are measured.

Table A4.2: Structure of Saarland data

Academic year	Enrollment in grade 1 (School-level)	Test data in grade 3 (Student-level)
2000/01	✓	
2001/02	✓	
2002/03	✓	
2003/04	✓	✓
2004/05	✓	✓
2005/06	✓	✓
2006/07		✓

Notes: Enrollment refers to data on the number of students in grade 1 in the respective academic year who were enrolled one year late, enrolled one year early, and retained in the previous year. Source: Statistisches Amt des Saarlands (2017); State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.3: Structure of NEPS data

	2011 Wave 1	2012 Wave 2	2013 Wave 3	2013/2014 Wave 4	2014/2015 Wave 5	2015/2016 Wave 6
			Expected Grade:			
			1	2	3	4
Language						
Reading Competence				✓		✓
Reading Speed			✓			
Vocabulary			✓		✓	
Grammar			✓			
Math			✓	✓		✓
Cognition				✓		

Notes: The expected grade refers to the grade that a student should be in if (s)he was enrolled on time and did not skip or repeat a grade. Source: NEPS Data, Data Version SC2: 6.0.1.

Table A4.4: Effects of cohort size on the grade-level student composition for Saxony

	% Late enrolled	% Early enrolled	% Repeater		
	Grade 1			Grade 2	Grade 3
	(1)	(2)	(3)	(4)	(5)
	Panel A: OLS grade composition				
Imputed cohort size	-0.048**	-0.011***	-0.048***	-0.058**	-0.074**
	(0.024)	(0.004)	(0.016)	(0.024)	(0.031)
	Panel B: IV grade composition				
Class size	-0.495***	-0.070***	-0.362***	-0.602***	-1.036***
	(0.044)	(0.015)	(0.026)	(0.044)	(0.082)
N SchoolYearObs	3921	3921	3921	3921	3921

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each cell contains results for separate, weighted regression with weights equal to total enrollment. Columns 1-3 in panel A report estimates of the effects of imputed cohort size on the percentage of repeating-, late- and early enrolled students in grade 1. Columns 4-5 report estimates of the effects of imputed cohort size on the percentage of repeating students in grade 2 and grade 3, respectively. Columns 1-3 in panel B report instrumental variables estimates of average class size in grade 1 on the percentage of repeating-, late- and early enrolled students in grade 1. The instrument for class size is imputed cohort size divided by the number of classes. Columns 4-5 report instrumental variables estimates of average class size in grades 2 and 3 on the percentage of repeating-, late- and early enrolled students in grades 2 and 3. The instrument for class size in the respective grade is imputed cohort size divided by the number of classes. Regressions include school and year fixed effects. Standard errors clustered at the school-level are given in parentheses. Source: Statistisches Landesamt Sachsen (2017).

Table A4.5: Full results: The effect of class size on language test scores

	OLS				IV			
	Avg. class size grade 3				IV: Imputed cohort size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AvgclassSizeGrade3	-0.016***	-0.018***	-0.020***	-0.020***	-0.007	-0.015*	-0.019**	-0.019**
	(0.004)	(0.004)	(0.005)	(0.005)	(0.009)	(0.009)	(0.009)	(0.009)
2004.year	-0.003	0.002	0.001	-0.458***	-0.001	0.003	0.001	-0.457***
	(0.025)	(0.024)	(0.026)	(0.054)	(0.025)	(0.024)	(0.026)	(0.054)
2005.year	0.016	-0.020	-0.155***	-0.607***	0.004	-0.024	-0.157***	-0.608***
	(0.035)	(0.035)	(0.045)	(0.061)	(0.036)	(0.036)	(0.047)	(0.063)
2006.year	0.004	-0.025	-0.157***	-0.574***	-0.005	-0.028	-0.158***	-0.575***
	(0.033)	(0.033)	(0.040)	(0.057)	(0.033)	(0.034)	(0.041)	(0.058)
9.ageIM	—	-0.126***	-0.088***	-0.065***	—	-0.126***	-0.088***	-0.065***
		(0.014)	(0.013)	(0.013)		(0.014)	(0.013)	(0.013)
10.ageIM	—	-0.881***	-0.584***	-0.517***	—	-0.881***	-0.584***	-0.517***
		(0.025)	(0.023)	(0.022)		(0.025)	(0.023)	(0.022)
11.ageIM	—	-1.156***	-0.757***	-0.642***	—	-1.156***	-0.757***	-0.642***
		(0.051)	(0.047)	(0.046)		(0.051)	(0.047)	(0.046)
99.ageIM	—	-0.431***	-0.367***	-0.149	—	-0.432***	-0.367***	-0.149
		(0.102)	(0.112)	(0.209)		(0.103)	(0.112)	(0.209)
5.germanIM	—	—	-0.909***	-0.833***	—	—	-0.909***	-0.833***
			(0.016)	(0.015)			(0.016)	(0.015)
99.germanIM	—	—	-0.389***	-0.373***	—	—	-0.389***	-0.373***
			(0.047)	(0.046)			(0.047)	(0.046)
1.maleIM	—	—	—	-0.136***	—	—	—	-0.136***
				(0.009)				(0.009)
3.maleIM	—	—	—	-0.194	—	—	—	-0.194
				(0.179)				(0.179)
1.booksIM	—	—	—	0.206***	—	—	—	0.206***
				(0.028)				(0.028)
2.booksIM	—	—	—	0.341***	—	—	—	0.341***
				(0.026)				(0.026)
3.booksIM	—	—	—	0.406***	—	—	—	0.406***
				(0.026)				(0.026)
4.booksIM	—	—	—	0.476***	—	—	—	0.476***
				(0.028)				(0.028)
5.booksIM	—	—	—	-0.110**	—	—	—	-0.110**
				(0.054)				(0.054)
1.migIM	—	—	—	-0.059	—	—	—	-0.059
				(0.037)				(0.037)
2.migIM	—	—	—	-0.194**	—	—	—	-0.195**
				(0.076)				(0.077)
1.foreign	—	—	—	-0.076**	—	—	—	-0.076**
				(0.032)				(0.032)
2.foreign	—	—	—	0.107	—	—	—	0.108
				(0.093)				(0.094)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Durbin-Wu-Hausman test statistic					1.485	0.227	0.028	0.011
P-Value Durbin-Wu-Hausman test					0.223	0.633	0.868	0.918
N	37847	37847	37847	37847	37847	37847	37847	37847

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each column contains results for a separate regression. Columns 1-4 report estimates of class size in grade 3 on language. Columns 5-8 report estimates of class size in grade 3 where class size is instrumented by predicted class size based on imputed cohort size. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.6: Full results: The effect of class size on math test scores

	OLS				IV			
	Avg. class size grade 3				IV: Imputed cohort size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AvgclassSizeGrade3	-0.011	-0.013*	-0.014**	-0.014**	-0.006	-0.012	-0.015	-0.014
	(0.007)	(0.007)	(0.007)	(0.007)	(0.011)	(0.011)	(0.011)	(0.011)
2004.year	-0.003	0.000	-0.000	-0.321***	-0.002	0.000	-0.000	-0.321***
	(0.032)	(0.032)	(0.034)	(0.047)	(0.032)	(0.032)	(0.034)	(0.047)
2005.year	-0.027	-0.056	-0.153***	-0.468***	-0.034	-0.056	-0.152***	-0.468***
	(0.045)	(0.045)	(0.049)	(0.060)	(0.047)	(0.047)	(0.051)	(0.062)
2006.year	-0.037	-0.059	-0.154***	-0.442***	-0.042	-0.059	-0.154***	-0.442***
	(0.046)	(0.046)	(0.049)	(0.061)	(0.047)	(0.047)	(0.050)	(0.062)
9.ageIM	—	-0.079***	-0.051***	-0.052***	—	-0.079***	-0.051***	-0.052***
		(0.016)	(0.015)	(0.015)		(0.016)	(0.015)	(0.015)
10.ageIM	—	-0.691***	-0.472***	-0.455***	—	-0.691***	-0.472***	-0.455***
		(0.025)	(0.024)	(0.023)		(0.025)	(0.024)	(0.023)
11.ageIM	—	-0.842***	-0.551***	-0.515***	—	-0.842***	-0.551***	-0.515***
		(0.049)	(0.047)	(0.046)		(0.049)	(0.047)	(0.046)
99.ageIM	—	-0.328**	-0.309**	-0.004	—	-0.328***	-0.309**	-0.004
		(0.127)	(0.131)	(0.192)		(0.127)	(0.131)	(0.192)
5.germanIM	—	—	-0.668***	-0.654***	—	—	-0.668***	-0.654***
			(0.017)	(0.017)			(0.017)	(0.017)
99.germanIM	—	—	-0.254***	-0.237***	—	—	-0.254***	-0.237***
			(0.053)	(0.054)			(0.053)	(0.054)
1.maleIM	—	—	—	0.204***	—	—	—	0.204***
				(0.009)				(0.009)
3.maleIM	—	—	—	-0.140	—	—	—	-0.140
				(0.144)				(0.144)
1.booksIM	—	—	—	0.183***	—	—	—	0.183***
				(0.030)				(0.030)
2.booksIM	—	—	—	0.323***	—	—	—	0.323***
				(0.031)				(0.031)
3.booksIM	—	—	—	0.375***	—	—	—	0.375***
				(0.033)				(0.033)
4.booksIM	—	—	—	0.442***	—	—	—	0.442***
				(0.034)				(0.034)
5.booksIM	—	—	—	0.010	—	—	—	0.010
				(0.049)				(0.049)
1.migIM	—	—	—	0.024	—	—	—	0.024
				(0.044)				(0.043)
2.migIM	—	—	—	-0.116	—	—	—	-0.116
				(0.071)				(0.072)
1.foreign	—	—	—	0.005	—	—	—	0.005
				(0.038)				(0.037)
2.foreign	—	—	—	0.029	—	—	—	0.029
				(0.104)				(0.104)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Durbin-Wu-Hausman test statistic					0.309	0.005	0.006	0.000
P-Value Durbin-Wu-Hausman test					0.578	0.944	0.939	1.000
N	36845	36845	36845	36845	36845	36845	36845	36845

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each column contains results for a separate regression. Columns 1-4 report estimates of class size in grade 3 on math. Columns 5-8 report estimates of class size in grade 3 where class size is instrumented by predicted class size based on imputed cohort size. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.7: The effect of class size in different grades on test scores

	OLS				IV			
	Avg. class size in							
	Grade 1	Grade 2	Grade 3	Grade 1-3	Grade 1	Grade 2	Grade 3	Grade 1-3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Language	-0.0109**	-0.0105**	-0.0199***	-0.0153***	-0.0140**	-0.0171**	-0.0191**	-0.0160**
	(0.0055)	(0.0050)	(0.0050)	(0.0054)	(0.0068)	(0.0080)	(0.0092)	(0.0077)
Math	-0.0095	-0.0061	-0.0140**	-0.0109	-0.0102	-0.0123	-0.0140	-0.0117
	(0.0068)	(0.0067)	(0.0070)	(0.0074)	(0.0080)	(0.0095)	(0.0110)	(0.0092)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Limited German proficiency	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N Cluster	156	156	156	156	156	156	156	156
N SchoolYearObs	828	828	828	828	828	828	828	828

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each cell contains results for a separate regression. Columns 1-4 report estimates of class size in different grades on language and math. Columns 5-8 report estimates of class size in different grades where class size is instrumented by predicted class size based on imputed cohort size. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.8: Spline IV regressions

	17.5	18.5	19.5	20.5	21.5	22.5	23.5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Panel A: Language						
Class size $<$ knot	0.0798**	0.0373	0.0148	0.0006	-0.0146	-0.0214	-0.0230*
	(0.0397)	(0.0294)	(0.0242)	(0.0202)	(0.0168)	(0.0147)	(0.0134)
Class size $\geq$ knot	-0.0428***	-0.0424***	-0.0436***	-0.0458***	-0.0379	-0.0284	-0.0235
	(0.0119)	(0.0126)	(0.0141)	(0.0171)	(0.0232)	(0.0343)	(0.0549)
N	37847	37847	37847	37847	37847	37847	37847
Cragg-Donald Wald F statistic	5355	5446	5236	4600	3355	2087	1365
Kleibergen-Paap rk Wald F statistic	58.75	66.24	53.35	34.27	17.38	8.00	3.86
	Panel B: Math						
Class size $<$ knot	0.0943**	0.0484	0.0246	0.0150	-0.0054	-0.0185	-0.0249
	(0.0458)	(0.0332)	(0.0278)	(0.0238)	(0.0206)	(0.0183)	(0.0167)
Class size $\geq$ knot	-0.0390**	-0.0387**	-0.0405**	-0.0489**	-0.0390	-0.0148	0.0267
	(0.0153)	(0.0163)	(0.0189)	(0.0237)	(0.0323)	(0.0484)	(0.0765)
N	36845	36845	36845	36845	36845	36845	36845
Cragg-Donald Wald F statistic	5203	5293	5084	4465	3254	2009	1310
Kleibergen-Paap rk Wald F statistic	58.74	66.57	53.32	34.09	17.15	7.80	3.76
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Limited German Proficiency	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports IV results for different linear spline specifications where we instrument the linear spline in average class size in grade 3 by the linear spline in predicted class size based on imputed cohort size. All splines are estimated with one knot whose position is indicated in the column header. The coefficients measure class size effects for the specified interval. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.9: Robustness checks: Different specifications

	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
Language	-0.020***	-0.027***	-0.020***	-0.019**	-0.031	-0.016
	(0.005)	(0.010)	(0.007)	(0.009)	(0.020)	(0.015)
N	37847	15386	37847	37847	15386	37847
Cragg-Donald Wald F statistic				17017	4484	11648
Kleibergen-Paap rk Wald F statistic				176.48	38.42	86.29
Math	-0.014**	-0.019	-0.021**	-0.014	-0.041	-0.021
	(0.007)	(0.012)	(0.009)	(0.011)	(0.026)	(0.018)
N	36845	14944	36845	36845	14944	36845
Cragg-Donald Wald F statistic				16614	4366	11304
Kleibergen-Paap rk Wald F statistic				175.77	38.05	84.89
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes	Yes	Yes
Limited German proficiency	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes			Yes	Yes
School-specific linear trends		Yes			Yes	
School-number of classes combination FE			Yes			Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. Each cell contains results for a separate regression. Columns 1-4 report estimates of class size in grade 3 on language and math. Columns 5-8 report estimates of class size in grade 3 where class size is instrumented by predicted class size based on imputed cohort size. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.10: Robustness checks: Different linear spline regressions with knot at class size 20.5

	OLS			IV		
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: Language			
Class size $<$ knot	0.007	0.005	-0.003	0.001	-0.007	0.017
	(0.010)	(0.019)	(0.013)	(0.021)	(0.061)	(0.029)
Class size $\geq$ knot	-0.041***	-0.045***	-0.034***	-0.039**	-0.048	-0.057*
	(0.009)	(0.017)	(0.011)	(0.017)	(0.045)	(0.032)
N	37847	15386	37847	37847	15386	37847
Cragg-Donald Wald F statistic				4300	745	2270
Kleibergen-Paap rk Wald F statistic				32.41	6.32	14.63
			Panel A: Math			
Class size $<$ knot	0.008	0.020	-0.001	0.014	0.062	0.042
	(0.012)	(0.027)	(0.016)	(0.024)	(0.069)	(0.036)
Class size $\geq$ knot	-0.031**	-0.041*	-0.038**	-0.042*	-0.111*	-0.101**
	(0.013)	(0.021)	(0.016)	(0.024)	(0.060)	(0.042)
N	36845	14944	36845	36845	14944	36845
Cragg-Donald Wald F statistic				4174	716	2207
Kleibergen-Paap rk Wald F statistic				32.27	6.18	14.33
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes	Yes	Yes
Limited German proficiency	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes			Yes	Yes
School specific linear trends		Yes			Yes	
School-number of classes combination FE			Yes			Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports IV results for different linear spline specifications for class size in grade 3 with a single knot at 20.5 . The coefficients measure class size effects for the specified interval. Columns 1-4 report OLS results. Columns 5-8 report estimates where we instrument the linear spline in class size in grade 3 by a linear spline in predicted class size in based on imputed cohort size. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include age in years, gender, number of books at home, migration background, and native language for regressions on language and math test scores. The regressions on the migrant share do not include individual control variables. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.11: Heterogeneity IV

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Panel A: Language			
Avg. class size grade 3	-0.019**	-0.018*	-0.018*	-0.017*	-0.018*	-0.017*	-0.017*
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
× female	0.000						
	(0.004)						
× older than 9 years		-0.011					
		(0.009)					
× few books			-0.011				
			(0.007)				
× migration background				-0.019**			
				(0.008)			
× insufficient German proficiency					-0.035***		
					(0.001)		
× dyslexia						-0.041***	
						(0.001)	
× dyscalculia							-0.032***
							(0.001)
N	37847	37847	37847	37847	37847	37847	37847
Cragg-Donald Wald F statistic	8502	8481	8422	8338	8509	8508	8510
Kleibergen-Paap rk Wald F statistic	88.43	88.25	89.39	87.55	88.24	88.24	88.30
				Panel B: Math			
Avg. class size grade 3	-0.011	-0.012	-0.013	-0.013	-0.013	-0.013	-0.012
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
× female	-0.006						
	(0.005)						
× older than 9 years		-0.018*					
		(0.010)					
× few books			-0.011				
			(0.007)				
× migration background				-0.010			
				(0.008)			
× insufficient German proficiency					-0.024***		
					(0.001)		
× dyslexia						-0.023***	
						(0.001)	
× dyscalculia							-0.044***
							(0.001)
N	36845	36845	36845	36845	36845	36845	36845
Cragg-Donald Wald F statistic	8300	8285	8217	8114	8308	8307	8308
Kleibergen-Paap rk Wald F statistic	88.12	87.78	89.03	87.09	87.89	87.88	87.95
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Limited German proficiency	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. This table reports IV results where each column in panels A and B contains the results of a separate regression with the same specification as in column 6 of Table 4.8, except that the class size variable is interacted with an indicator variable for the individual student characteristics. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Individual controls include age in years, gender, number of books at home, migration background, and native language. Source: State-wide Orientation Exams 2003-2006, see Paulus and Leidinger (2009).

Table A4.12: Monte Carlo simulation

	Balancing	Reduced form	IV
	(1)	(2)	(3)
	Panel A: Grade 1		
Mean $\hat{\beta}$	0.001	-0.057	-0.267
Mean SE of $\hat{\beta}$	0.043	0.010	0.010
95% Lower Bound	-0.019	-0.077	-0.352
95% Upper Bound	0.019	-0.038	-0.187
	Panel B: Grade 2		
Mean $\hat{\beta}$	-0.000	-0.105	-0.404
Mean SE of $\hat{\beta}$	0.084	0.009	0.013
95% Lower Bound	-0.018	-0.129	-0.592
95% Upper Bound	0.018	-0.082	-0.253
	Panel C: Grade 3		
Mean $\hat{\beta}$	0.000	-0.149	-0.507
Mean SE of $\hat{\beta}$	0.121	0.009	0.015
95% Lower Bound	-0.018	-0.177	-0.766
95% Upper Bound	0.019	-0.122	-0.277

Notes: 1000 iterations, 95% confidence bounds are obtained from 25th and 975th estimate of ordered $\hat{\beta}$.

Appendix B: Data

B4.1 State-wide Orientation Exams Saarland

For 2003 and 2004, the development of test items for the centralized exams was carried out by the Bavarian State Institute of School Quality and Education Research, an organization with more than 50 years of experience in the field of educational consulting. In 2005 and 2006, this responsibility was transferred to Saarland's standing conferences on language and mathematics (Landesfachkonferenzen). Since the aim of the SOE was to safeguard quality assurance, test items were created such that they could assess students' competences in relation to education standards set by the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder (Kultusministerkonferenz). The subject matter of the tests was the material from grades 2 and 3. In German, this related to the two domains of "Reading" and "Writing / Language and Use of Language." In reading, reference was made to the cognitive model of van Dijk and Kintsch (1983) that is also used in the international PIRLS studies. Questions were multiple choice and required extracting pieces of information from short texts. The most difficult questions further entailed meta-cognitive abilities, for example in the sense of relating texts to the author's likely intentions of writing them. In the domain of writing and use of language, spelling and grammar competences were specifically tested. Therefore, students had to complete words and reformulate sentences. The mathematics test was not further subdivided into different domains. However, all questions pertained to one or more of the following general mathematical competences: modelling, problem solving, argumentation, illustration, and communication. These competences had to be applied to specific mathematical content that students were supposed to be familiar with (Paulus and Leidinger, 2009).

B4.2 NEPS

The German National Education Panel Study (NEPS) was initially developed in 2009 to provide information on the determinants of education, the consequences of education, and to describe educational trajectories over the life course (Blossfeld et al., 2011). We use data from Starting Cohort 2, which is a nationwide, representative sample of children who were first surveyed as 4-year-olds in kindergarten in 2010/2011 and who were expected to begin schooling in the school year of 2012/2013.[122] We use data from Waves 3-6 during the academic years 2013/14-2015/2016, when these children should have been enrolled in grades 1-4. The NEPS interviews the children and parents separately. From the parents we know the year and month when a child first entered primary school and if a child repeated or skipped a grade. The NEPS provides standardized test scores to assess children's competencies in different dimensions. We compute language, math and cognition test scores by averaging the respective standardized test scores for each domain. For each respective score, Table A4.3 shows when each test was conducted that enters into each respective score. The cognition score is the average of standardized test scores of perceptual speed assessed by the Picture Symbol Test and reasoning assessed by matrices test.[123]

[122] For more information on the target population see Aßmann et al. (2011).

[123] The Picture Symbol Test is based on an improved version of the Digit-Symbol Test (DST) from the tests of the Wechsler family by Lang et al. (2007). Each item of the matrices test for reasoning consists of several horizontally and vertically arranged fields in which different geometrical elements are shown with only one field remaining free. The logical rules on which the pattern of the geometrical elements is based have to be deduced in order to be able to select the right complement for the free field from the offered solutions.

Appendix C: Model Extensions

C4.1 School System with Redshirting

Modifying our model to allow for redshirting corresponds to a simple relabeling of our model in section 4.3. LG now refers to the years in child care before school entry and HG to the first grade in primary school. Children spend L years in child care. The grade retention threshold p is the academic skill level that children must attain to be enrolled in first grade. Children with academic skills below this threshold spend another year in child care, thus entering grade 1 a year later. λ_s^t is equal to the share of students from birth cohort t who enter grade 1 (HG) without being redshirted and ϕ_s^τ is equal to the share of children in grade 1 in year τ who were enrolled on schedule. π^α and π^ω capture the effects of class size on academic skills in child care and grade 1, respectively. The average test performance of students who were enrolled on time is then given in Equation (4.10) and the average test performance of redshirted students is given in Equation (4.11), where δ_s^t captures school and birth cohort-specific changes in skills associated with redshirting.

C4.2 School System with Early Enrollment

To allow for early school enrollment in our model in section 4.3, we apply the same relabeling as in the model with redshirting. The only difference to the model with redshirting is that if children attain the threshold p, they are enrolled in first grade one year earlier than regular students (after $L-1$ instead of L years). Following the line of reasoning in section 4.3, the share of students from birth cohort t who enter grade 1 (HG) regularly in year $t+L$ is

$$\lambda_s^t = \frac{-\alpha_s^t + \theta + p_s^t}{2\theta} \tag{C4.1}$$

Class size in HG in school s in the school year starting in τ depends on the size of cohorts $\tau - L$ and $\tau - L + 1$ as well as the share of regularly enrolled students in these birth cohorts

$$N_{s\tau}^{obs} = \lambda_s^{\tau-L} N_s^{\tau-L} + (1-\lambda_s^{\tau-L+1}) N_s^{\tau-L+1} \tag{C4.2}$$

The share of regularly enrolled students in HG in school s in the school year starting in τ is then

$$\phi_s^\tau = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{N_{s\tau}^{obs}} = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{\lambda_s^{\tau-L} N_s^{\tau-L} + (1-\lambda_s^{\tau-L+1}) N_s^{\tau-L+1}} \tag{C4.3}$$

Students take a standardized test at the end of HG. The test performance of regularly enrolled students reflects their academic skills accumulated in LG and HG, $a_{is}^t + \omega_{s,t+L}$. The average test performance of these students from cohort t who reach HG in year $\tau = t + L$ can be written as

$$E\left(test_{is}^t | regular\right) = E\left(test_{is}^t | test_{is}^t < p_s^t\right) = \frac{\alpha_s^t - \theta + p_s^t}{2} + \omega_{s,t+L} \tag{C4.4}$$

where $\omega_{s,t+L}$ denotes the average skills these students accumulate in HG in year $t + L$. The test performance of early enrolled students who reach HG one year earlier is $a_{is}^t + w_{s,t+L+1} + \delta_s^t$, where δ_s^t captures a school and birth cohort-specific change in skills associated with early enrollment. This change in skills may be positive or negative. The average performance of these early enrolled students in HG is

$$E\left(test_{is}^t | early\right) = E\left(test_{is}^t | test_{is}^t \geq p_s^t\right) = \frac{\alpha_s^t + \theta + p_s^t}{2} + \delta_s^t + \omega_{s,t+L-1} \tag{C4.5}$$

The average test performance of all students in HG in year τ is then

$$test_{s\tau} = \phi_s^{\tau-L} E\left(test_{is}^{\tau-L} | regular\right) + (1-\phi_s^{\tau-L}) E\left(test_{is}^{\tau-L+1} | early\right) \tag{C4.6}$$

Appendix D: Proofs

To prove the results in section 4.3, note that in the case of two periods, the within-school estimator is equivalent to the first difference estimator. We first linearize the within-school change in observed class size in high grade (HG), $\Delta N^{obs}_{s\tau} = N^{obs}_{s\tau} - N^{obs}_{s,\tau-1}$, around $N^t_s = N$, $\alpha^t_s = \alpha$, and $p^t_s = p$ and we assume w.l.o.g. that $N = 1$. Making use of Equation (4.6) and Equation (4.7), this yields

$$\begin{aligned}\Delta N^{obs}_{s\tau} = &\left(\frac{\pi^\alpha}{2\theta} + \lambda\right)\Delta N_s^{\tau-L} + \left(1 - \lambda - \frac{\pi^\alpha}{2\theta}\right)\Delta N_s^{\tau-L-1} \\ &+ \frac{1}{2\theta}\left(\Delta\alpha_s^{\tau-L} - \Delta\alpha_s^{\tau-L-1} - \Delta p_s^{\tau-L} + \Delta p_s^{\tau-L-1}\right)\end{aligned} \tag{D4.1}$$

where $\lambda = \frac{\alpha+\theta+p}{2\theta}$, $\Delta N^t_s = N^t_s - N^{t-1}_s$, $\Delta\alpha^t_s = \alpha^t_s - \alpha^{t-1}_s$ and $\Delta p^t_s = p^t_s - p^{t-1}_s$. Linearizing the within-school change in the average test score in HG, $\Delta test_{s\tau} = test_{s\tau} - test_{s,\tau-1}$, using (4.2)-(4.12) yields

$$\begin{aligned}\Delta test_{s\tau} = &\left[\left(\lambda + \frac{\pi^\alpha}{2\theta}\right)(1-\lambda)(\theta-\delta) + \lambda\frac{\pi^\alpha}{2} + \pi^\omega(\lambda + \frac{\pi^\alpha}{2\theta})\right]\Delta N_s^{\tau-L} \\ &+ \left[\lambda\left(\frac{\pi^\alpha}{2\theta} - 1 + \lambda\right)(\theta-\delta) + \frac{\pi^\alpha}{2}(1-\lambda) + \pi^\omega(1 - \lambda - \frac{\pi^\alpha}{2\theta})\right]\Delta N_s^{\tau-L-1} \\ &+ \left((\theta-\delta)\frac{1-\lambda}{2\theta} + \frac{\lambda}{2}\right)\left(\Delta\alpha_s^{\tau-L} - \Delta p_s^{\tau-L}\right) \\ &+ \left((\theta-\delta)\frac{\lambda}{2\theta} + \frac{1-\lambda}{2}\right)\left(\Delta\alpha_s^{\tau-L-1} - \Delta p_s^{\tau-L-1}\right)\end{aligned} \tag{D4.2}$$

D4.1 Retention Bias Without "True Class Size Effects"

To prove the result in (4.13), we assume that there are no class size effects, $\pi^\alpha = \pi^\omega = 0$, and that academic skills and the thresholds for grade retention are the same across schools and cohorts, $\alpha^t_s = \alpha$ and $p^t_s = p$. There are only shocks to cohort size as modeled in

Equation (4.2). In this case Equation (D4.1) and Equation (D4.2) simplify to

$$\Delta N^{obs}_{s\tau} = \lambda \Delta N^{\tau-L}_{s} + (1-\lambda)\,\Delta N^{\tau-L-1}_{s} \tag{D4.3}$$

$$\Delta test_{s\tau} = \lambda(1-\lambda)(\theta-\delta)\left(\Delta N^{\tau-L}_{s} - \Delta N^{\tau-L-1}_{s}\right) \tag{D4.4}$$

and the assumption of i.i.d. shocks to cohort size implies

$$\begin{aligned} Cov(\Delta test_{s\tau}, \Delta N^{\tau-L}_{s}) &= 3Var(\eta)(\theta-\delta)(1-\lambda)\lambda \\ Cov(\Delta N^{obs}_{s\tau}, \Delta N^{\tau-L}_{s}) &= Var(\eta)(3\lambda-1) \end{aligned} \tag{D4.5}$$

The IV estimate is equal to the ratio of these two covariances

$$\begin{aligned} \beta_{IV} &= \frac{Cov(\Delta test_{s\tau}, \Delta N^{\tau-L}_{s})}{Cov(\Delta N^{obs}_{s\tau}, \Delta N^{\tau-L}_{s})} \\ &= \frac{3(\theta-\delta)(1-\lambda)\lambda}{3\lambda-1} \end{aligned} \tag{D4.6}$$

which is positive if students retained in the past perform on average worse than non-retained students, $\theta - \delta > 0$, and less than $2/3$ of all students are retained ($\lambda > 1/3$).

D4.2 IV Results

To derive β_{IV} in Equation (4.14), we need to calculate the covariances $Cov(\Delta test_{s,\tau}, \Delta N^{obs}_{s\tau})$ and $Cov(\Delta N^{obs}_{s\tau}, \Delta N^{\tau-1}_{s})$. Under our assumption of i.i.d. shocks to the cohort size N^{t}_{s}, η^{t}_{s}, it is straightforward to show

$$Cov(\Delta N^{obs}_{s\tau}, \Delta N^{\tau-L}_{s}) = Var(\eta)\left(3\frac{\pi^{\alpha}}{2\theta} + 3\lambda - 1\right) \tag{D4.7}$$

and

$$\begin{aligned} Cov(\Delta test^{obs}_{s\tau}, \Delta N^{\tau-L}_{s}) &= Var(\eta)(\theta-\delta)\left[3\lambda(1-\lambda) + \frac{\pi^{\alpha}}{2\theta}(2-3\lambda)\right] \\ &+ Var(\eta)\left[\frac{\pi^{\alpha}}{2}(3\lambda-1) + \pi^{\omega}\left(3\frac{\pi^{\alpha}}{2\theta} + 3\lambda - 1\right)\right] \end{aligned} \tag{D4.8}$$

Taking the ratio of (D4.8) and (D4.7) gives the IV estimate

$$\begin{aligned} \beta_{IV} &= \frac{Cov(\Delta test_{s,\tau}, \Delta N_s^{\tau-L})}{Cov(\Delta N_{s\tau}^{obs}, \Delta N_s^{\tau-L})} \\ &= \rho_{IV}\,(\theta - \delta) + \xi_{IV}\pi^{\alpha} + \pi^{\omega} \end{aligned} \tag{D4.9}$$

where

$$\rho_{IV} = \frac{3\lambda(1-\lambda) + \frac{\pi^{\alpha}}{2\theta}(2-3\lambda)}{3\frac{\pi^{\alpha}}{2\theta} + 3\lambda - 1} \tag{D4.10}$$

and

$$\xi_{IV} = \frac{1}{2}\frac{3\lambda - 1}{3\frac{\pi^{\alpha}}{2\theta} + 3\lambda - 1} \tag{D4.11}$$

ξ_{IV} will be approximately equal to 1/2. To see this note that $-\pi^{\alpha}/2\theta$ is the marginal effect of class size in LG on the share of grade repeaters in LG.[124] This effect is likely to be very small relative to $3\lambda - 1$ and therefore can be neglected.[125] Using the same argument, it is easy to see from Equation (D4.10) that $\rho_{IV} \geq 0$ if class size has a negative effect on skills in LG, $\pi^{\alpha} < 0$, and the share of retained students is smaller than $1/3$.

Analogous arguments yield that the terms in Equation (D4.10), which include $\pi^{\alpha}/2\theta$, have only a negligible impact on the size of ρ_{IV}.

D4.2.1 IV Result Controlling for the Effect of Grade Retention at the Individual Level

To derive β_{IV}^{REA} in Equation (4.15) for the instrumental-variables approach, notice that controlling for the effect of grade retention on academic achievement at the individual level is equivalent to adjusting the academic achievement of retained students by the average gap in academic achievement between retained and non-retained students in the same grade and school. This gap is $\theta - \delta$, see Equation (4.10) and Equation (4.11). Therefore,

[124] To see this, simply take the derivative of $1 - \lambda_s^t$ with respect to N_s^t using Equation (4.6).

[125] Our estimate for the marginal effect of class size on the share of grade repeaters in grade 1 is 0.0015 (see column 4 of Table 4.11). If we assume this effect is constant for grades 1 through 3, this estimate implies a value of $\pi^{\omega}/2\theta$ equal to 0.0045. Multiplying this by 3 still gives a value that is two orders of magnitude smaller than our estimate for $3\lambda - 1$, which is equal to 1.67 given that the average accumulated retention rate in grade 3 ($= 1 - \lambda$ in our setting) is equal to 0.11 (see Table 4.2).

the average test score in HG adjusted for the effect of grade retention at the individual level becomes

$$test_{s\tau}^{REA} = \phi_s^\tau E\left(test_{is}^\tau | non - retained\right) + \left(1 - \phi_s^\tau\right)\Big(E\left(test_{is}^\tau | retained\right) + (\theta - \delta)\Big) \tag{D4.12}$$

which differs from $test_{s\tau}$ in Equation (4.12) only in the $\theta - \delta$ term. Linearizing $\Delta test_{s\tau}^{REA} = test_{s\tau}^{REA} - test_{s\tau-1}^{REA}$ by following the same steps we used to obtain Equation (D4.2) then yields

$$\begin{aligned} \Delta test_{s\tau}^{REA} &= \left[\lambda\frac{\pi^\alpha}{2} + \pi^\omega(\lambda + \frac{\pi^\alpha}{2\theta})\right]\Delta N_s^{\tau-L} \\ &+ \left[\frac{\pi^\alpha}{2}(1-\lambda) + \pi^\omega(1 - \lambda - \frac{\pi^\alpha}{2\theta})\right]\Delta N_s^{\tau-L-1} \\ &+ \frac{\lambda}{2}\left(\Delta\alpha_s^{\tau-L} - \Delta p_s^{\tau-L}\right) + \frac{1-\lambda}{2}\left(\Delta\alpha_s^{\tau-L-1} - \Delta p_s^{\tau-L-1}\right) \end{aligned} \tag{D4.13}$$

The covariance of $\Delta test_{s\tau}^{REA}$ and $\Delta N_s^{\tau-L}$ can be shown to be

$$Cov(\Delta test_{s\tau}^{REA}, \Delta N_s^{\tau-L}) = Var(\eta)\left[\frac{\pi^\alpha}{2}(3\lambda - 1) + \pi^\omega\left(3\frac{\pi^\alpha}{2\theta} + 3\lambda - 1\right)\right] \tag{D4.14}$$

Taking the ratio of (D4.14) and (D4.7) gives the IV estimate when controlling for grade retention on the individual level

$$\begin{aligned} \beta_{IV}^{REA} &= \frac{Cov(\Delta test_{s,\tau}^{REA}, \Delta N_s^{\tau-L})}{Cov(\Delta N_{s\tau}^{obs}, \Delta N_s^{\tau-L})} \\ &= \xi_{IV}\pi^\alpha + \pi^\omega \end{aligned} \tag{D4.15}$$

where ξ_{IV} is defined in Equation (D4.11).

D4.3 OLS Results

To derive β_{OLS} in Equation (4.16), we need to calculate the variance of $\Delta N_{s\tau}^{obs}$ and the covariance of $\Delta test_{s,\tau}$ and $\Delta N_{s\tau}^{obs}$. Under our assumption of i.i.d. shocks to N_s^t, α_s^t, and

p_s^t it is straightforward to show that

$$\begin{aligned} Var(\Delta N_{s\tau}^{obs}) = 2Var(\eta) & \left((\lambda + \frac{\pi^\alpha}{2\theta})^2 + (1 - \lambda - \frac{\pi^\alpha}{2\theta})^2 - (\lambda + \frac{\pi^\alpha}{2\theta})(1 - \lambda - \frac{\pi^\alpha}{2\theta}) \right) \\ & + \frac{6}{4\theta^2} \left(Var(\epsilon) + Var(\nu) \right) \end{aligned} \tag{D4.16}$$

and

$$\begin{aligned} Cov(\Delta test_{s\tau}, \Delta N_{s\tau}^{obs}) = (\theta - \delta) & \Bigg[Var(\eta) \left((\lambda + \frac{\pi^\alpha}{2\theta})(1 - \lambda) \left(\lambda + \lambda^2 + \frac{\pi^\alpha}{2\theta}(2 + \lambda) \right) \right. \\ & \left. + \lambda(1 - \lambda - \frac{\pi^\alpha}{2\theta}) \left(3\lambda + 3\frac{\pi^\alpha}{2\theta} - 2 \right) \right) \\ & + \left(Var(\epsilon) - Var(\nu) \right) \frac{1 - 2\lambda}{4\theta^2} \Bigg] \\ & + \left(Var(\epsilon) - Var(\nu) \right) \frac{6\lambda - 3}{4\theta} \\ & + \frac{\pi^\alpha}{2} Var(\eta)(2\lambda - 1) \left((3\lambda - 1)(\lambda + \frac{\pi^\alpha}{2\theta}) - (3\lambda - 2)(1 - \lambda - \frac{\pi^\alpha}{2\theta}) \right) \\ & + 2\pi^\omega Var(\eta) \left((\lambda + \frac{\pi^\alpha}{2\theta})^2 + (1 - \lambda - \frac{\pi^\alpha}{2\theta})^2 - (\lambda + \frac{\pi^\alpha}{2\theta})(1 - \lambda - \frac{\pi^\alpha}{2\theta}) \right) \end{aligned} \tag{D4.17}$$

Taking the ratio of (D4.17) and (D4.16) and collecting terms gives the OLS estimate

$$\begin{aligned} \beta_{OLS} & = \frac{Cov(\Delta test_{s,\tau}, \Delta N_{s\tau}^{obs})}{Var(\Delta N_{s\tau}^{obs})} \\ & = \rho_{OLS} \left(\theta - \delta \right) + \iota_{OLS} + \xi_{OLS} \pi^\alpha + \pi^\omega \end{aligned} \tag{D4.18}$$

where

$$\rho_{OLS} = \frac{ \begin{aligned} Var(\eta) \Big[(\lambda + \tfrac{\pi^\alpha}{2\theta})(1 - \lambda) \left(\lambda + \lambda^2 + \tfrac{\pi^\alpha}{2\theta}(2 + \lambda) \right) + \lambda(1 - \lambda - \tfrac{\pi^\alpha}{2\theta}) \left(3\lambda + 3\tfrac{\pi^\alpha}{2\theta} - 2 \right) \Big] \\ + \big(Var(\epsilon) - Var(\nu) \big) \tfrac{2\lambda - 1}{4\theta^2} \end{aligned} }{ Var(N_{s\tau}^{obs}) } \tag{D4.19}$$

and

$$\iota_{OLS} = \frac{(Var(\epsilon) - Var(\nu))\,\frac{6\lambda-3}{4\theta} - \pi^{\omega}\frac{6}{4\theta^2}\,(Var(\epsilon) + Var(\nu))}{Var(N^{obs}_{s\tau})} \tag{D4.20}$$

and

$$\xi_{OLS} = \frac{1}{2}\frac{Var(\eta)(2\lambda-1)\Big[(3\lambda-1)(\lambda+\frac{\pi^{\alpha}}{2\theta}) - (3\lambda-2)(1-\lambda-\frac{\pi^{\alpha}}{2\theta})\Big]}{Var(N^{obs}_{s\tau})} \tag{D4.21}$$

Using similar arguments about the relative magnitude of $\pi^{\alpha}/2\theta$ and λ as above, suggests that the terms involving $\pi^{\alpha}/2\theta$ in Equation (D4.19) and Equation (D4.21) can be neglected. In that case, it is easy to show that ($\xi_{OLS} < 1$). The signs of Equation (D4.19) and Equation (D4.20), however, depend on the difference in the variance of the shocks to ability levels and retention thresholds ($Var(\epsilon) - Var(\nu)$). Unless we make assumptions about the relative magnitudes of these shocks, the signs of ρ_{OLS} and ι_{OLS} are indeterminate.

D4.3.1 OLS Result Controlling for the Effect of Grade Retention at the Individual Level

Next, we derive β^{REA}_{OLS} in Equation (4.17) following the same logic as in the previous two sections. The covariance of $\Delta test^{REA}_{s\tau}$ and $\Delta N^{obs}_{s\tau}$ can be shown to be

$$\begin{aligned} Cov(\Delta test^{REA}_{s\tau}, \Delta N^{obs}_{s\tau}) = {} & (Var(\epsilon) - Var(\nu))\left[3\frac{2\lambda-1}{4\theta^2}\delta + 6\frac{\pi^{\omega}}{4\theta^2}\right] \\ & + Var(\eta)\left\{\frac{\pi^{\alpha}}{2}\left[4\lambda\frac{\pi^{\alpha}}{2\theta} - \frac{\pi^{\alpha}}{2\theta} + 4\lambda^2 - 2\lambda\right]\right. \\ & \left. + \pi^{\omega}\left[6\left(\frac{\pi^{\alpha}}{2\theta}\right)^2 - 6\frac{\pi^{\alpha}}{2\theta} - 12\lambda\frac{\pi^{\alpha}}{2\theta} + 6\lambda^2 - 6\lambda + 2\right]\right\} \end{aligned} \tag{D4.22}$$

Taking the ratio of (D4.22) and (D4.16) gives the OLS estimate with grade retention controls

$$\begin{aligned}\beta_{OLS}^{REA} &= \frac{Cov(\Delta test_{s,\tau}^{REA}, \Delta N_{s\tau}^{obs})}{Var(\Delta N_{s\tau}^{obs}} \\ &= \iota_{OLS} + \xi_{OLS}\pi^{\alpha} + \pi^{\omega}\end{aligned} \tag{D4.23}$$

where ι_{OLS} and ξ_{OLS} are defined in (D4.20) and (D4.21), respectively.

D4.4 Proofs for the Non-i.i.d. Case of Birth Cohort Size Shocks

In results, which we do not report here, we calculated autocorrelations for residuals from a regression of imputed cohort size on school-fixed effects. We find that these residuals have negative first- and second-order autocorrelations. This is consistent with the notion that women who give birth in year t are less likely to give birth in year $t + 1$ and $t + 2$. Thus, we investigate the implications of negatively autocorrelated shocks to the size of birth cohorts for the simple spurious class size effect without any "true class size effects". It can be shown that the spurious positive class size effect for the IV approach is even larger than in the i.i.d. case in Equation (4.13) under fairly general conditions. Theorem 1 summarizes this result: Let η_s^t be non-i.d.d. shocks that follow a stationary process. If

(i) less than one-third of all students are retained in LG ($\lambda \in (2/3, 1)$),

(ii) non-retained students have higher skills, on average, than students retained in the past ($\theta - \delta > 0$),

(iii) the first- and second order autocorrelations of η_s^t (ρ_1 and ρ_2) are negative but larger than -1 ($-1 < \rho_1, \rho_2 < 0$), and

(iv) the absolute value of the second-order autocorrelation of η_s^t is less than 3 times as large as the absolute value of its first-order autocorrelation ($3\rho_1 < \rho_2$),

then the IV approach in the absence of "true class size effects" yields a larger spurious positive class effect than in the i.d.d. case.

To prove Theorem 4, let ϕ_h denote the autocovariance of η_s^t between year t and $t + h$.

Using Equation (D4.3)-(D4.4) and stationarity of η_s^t yields

$$Cov\left(\Delta test_{s\tau}, \Delta N_s^{\tau-L}\right) = \lambda(1-\lambda)(\theta-\delta)\Big[3(\phi_0-\phi_1)+\phi_2\Big] \tag{D4.24}$$

$$Cov\left(\Delta N_{s\tau}^{obs}, \Delta N_s^{\tau-L}\right) = (3\lambda-1)\phi_0-(3\lambda-2)\phi_1+\lambda\phi_2 \tag{D4.25}$$

Taking the ratio of Equation (D4.24) and Equation (D4.25) yields the spurious class size effect for the case of non-i.i.d. shocks to birth cohort size

$$\frac{Cov\left(\Delta test_{s\tau}, \Delta N_s^{\tau-L}\right)}{Cov\left(\Delta N_{s\tau}^{obs}, \Delta N_s^{\tau-L}\right)} = \lambda(1-\lambda)(\theta-\delta)\frac{3(\phi_0-\phi_1)+\phi_2}{(3\lambda-1)\phi_0-(3\lambda-2)\phi_1+\lambda\phi_2} \tag{D4.26}$$

Let ρ_h denote the autocorrelation of η_t between time period t and $t+h$. In that case, expressing Equation (D4.26) in terms of autocorrelations yields

$$\lambda(1-\lambda)(\theta-\delta)\frac{3-3\rho_1+\rho_2}{(3\lambda-1)-(3\lambda-2)\rho_1+\lambda\rho_2} \tag{D4.27}$$

To complete the proof, it remains to be shown that Equation(D4.27) is greater than Equation (4.13) using conditions $(i)-(iv)$

$$\begin{aligned}\lambda(1-\lambda)(\theta-\delta)\frac{3-3\rho_1+\rho_2}{(3\lambda-1)-(3\lambda-2)\rho_1+\lambda\rho_2} &> \lambda(1-\lambda)(\theta-\delta)\frac{3-3\rho_1+\rho_2}{(3\lambda-2)+(3\lambda-2)\rho_1}\\ &> \lambda(1-\lambda)(\theta-\delta)\frac{3-3\rho_1+\rho_2}{2(3\lambda-2)}\\ &> \frac{3\lambda(1-\lambda)(\theta-\delta)}{2(3\lambda-2)}\\ &> \frac{3\lambda(1-\lambda)(\theta-\delta)}{(3\lambda-1)}\end{aligned}$$

Appendix E: Simulation

We test our theoretical predictions by running simulations of a school system that matches the school system in Saarland in terms of the average cohort size and the fraction of retained students in each grade. However, we abstract from the effect that class size has on retention rates and assume that the probability to be retained is constant across schools and cohorts. The data generating process is as follows:

- We create 268 primary schools. Each school s has an average cohort size in first grade equal to μ_s which is taken from a discrete uniform distribution with support $[20, 70]$.

- We then create 5 consecutive first-grade cohorts for each school, whose size is given by N_s^c, where c denotes the cohort.The N_s^c are random draws from a discrete uniform distribution with support $[0.8\mu_s, 1.2\mu_s]$. Thereby, we allow cohort size to fluctuate around the school's mean by 20%.

- Each student is retained at most once. The probabilities that a student is retained in first, second, or third grade are 3.2%, 2.9%, and 2.8%, respectively.

- We then create three grades for each cohort-school combination and assign students to each grade and cohort according to their retention status. For example, a student originally from cohort c who is retained in first grade is assigned to grade 1 of his initial cohort and to grades 1-3 of the next cohort ($c + 1$). The observed number of students in each school-grade-cohort is N_{scg}^{obs}, where g denotes the grade.

- In each grade, the number of classes is determined according to the class size rule:

$$C_{scg} = \frac{N_{scg}^{obs}}{int[(N_{scg}^{obs} - 1)/25] + 1} \tag{E4.1}$$

- Class size is equal to

$$CS_{scg} = \frac{N_{scg}^{obs}}{C_{scg}} \tag{E4.2}$$

- We drop the first cohort because it has no preceding cohort in which students can be retained.

We simulate the data 1,000 times and each time estimate three school-fixed-effects regressions separately for each grade: (1) we regress the fraction of students initially belonging to cohort c in grade 1 who are retained up to grade g on initial cohort size N_s^c; (2) we regress the fraction of students in grade g of cohort c who have previously been retained on the initial size of that cohort (N_s^c); (3) we regress the fraction of students in grade g of cohort c who have previously been retained on class size CS_{scg}, where we instrument class size by the predicted classes based on the initial cohort size (i.e. N_s^c/C_{scg}).

Descriptive statistics for the coefficients of cohort and class size from these estimations can be found in Table A4.12. By construction, belonging to an initially larger cohort (i.e. before cohort reassignment due to grade retention) is unrelated to whether or not a student will be retained. Hence, the coefficients for the initial cohort size in column 1 are close to zero. However, in column 2 we find a negative relationship between cohort size and the grade-level share of previously retained student in a cohort, which becomes stronger in higher grades. For the IV specification in column 2, we find a similar pattern with more than three times as large effects. Overall, the results for grade 1 are remarkably similar to those in column 3 of Table 4.3 based on actual data.

Chapter 5

Conclusion

This dissertation estimates cognitive skill returns to different features of education systems. It concentrates on policy interventions in young childhood, which is the age when children's general cognition is particularly malleable (see e.g. Cunha and Heckman, 2007; Hopkins and Bracht, 1975). The complexity of educational policy is acknowledged by an analysis of three very different policy levers. First, Chapter 2 focuses on early education by estimating the effect of an additional year of center-based child care attendance on a variety of educational outcomes, thereby estimating the effect of a *quantitative change* in education. On the contrary, Chapter 3 is dedicated to a *qualitative input* in education, namely teaching practices and their relation to achievement in primary school. Similarly, Chapter 4 deals with *another qualitative dimension* of education, namely the effects of class size changes on student achievement. The results of the three chapters show that cognitive skills can be increased in all three ways, either for all students or at least for certain subgroups. However, before jumping to (preliminary) conclusions for policy designs, one has to take a closer look at some of the limitations of the studies that may warrant more careful and nuanced interpretation of the results. As policy-makers are generally interested in the best available policy out of a larger set of feasible options, particular attention also has to be paid to expected effect sizes as well as efficiency issues, namely cost-benefit considerations. In the following, the results of this dissertation will be critically discussed against these benchmarks.

Chapter 2 investigates the question if longer child care attendance has positive effects on children in the medium- and long-run. The results suggest that cognitive skills as measured by German language grades at the age of 17 are significantly increased among treated children. Importantly, most of the total effect can be attributed to improved grades of children at the lower end of the achievement spectrum. Furthermore, treated children exhibit higher aspirations towards obtaining a vocational degree after high school as compared to not obtaining any further degree.

The main caveat of this study relates to what is effectively measured by the treatment. While it is true that treated children spend one more year in center-based child care than untreated children, they also enter child care at a younger age than other children. With the data at hand, it is impossible to disentangle a potential age-of-child-care-entry effect from the desired effect of longer child care attendance. Therefore, the results have to be interpreted as a combined effect of an additional year of child care attendance and a lower entry age. A further complication arises from the potentially important role played by peers. By entering child care earlier than their untreated counterparts, treated children were exposed to a mainly older peer group. Arguably, this first-hand learning experience from their more mature and able peers could explain some of their more favorable outcomes in adolescence. To what extent this is the case relates to the general question of the existence of peer effects in child care centers. The literature on this topic is relatively scarce. However, there is evidence for language peer effects that especially favor disadvantaged children (Justice et al., 2011). This finding underscores the possibility of young children learning from their older peers in precisely the domain of language development where large long-run effects are found in Chapter 2.

A further limitation of the study is its rather small sample size. This problem is especially salient in the preferred 2SLS specifications where only that part of the total variation in child care attendance is used that can be attributed to different regional supply levels. The small sample leads to imprecisely estimated treatment effects. When interpreting the effect sizes, this qualification has to be borne in mind. For that reason, it has been argued that rather than pinpointing exact effect sizes the results of this study should be understood as providing guidance on where to look for significant effects and what signs to expect. Any sophisticated cost-benefit analyses are therefore ruled out. The most modest reading of the results would in this context be that longer child care attendance is certainly not detrimental to child development. This in itself is an interesting finding, since the group of 'compliers' to which the results pertain consists of children who do not even stand to gain most from center-based child care attendance. These children who enter child care in an environment of excess demand come from families with a low resistance to center-based child care. One can assume that these families have on average rather advantaged social backgrounds since center-based child care take-up follows a social gradient in Germany (Bach et al., 2019; Cornelissen et al., 2018; Felfe and Lalive, 2018; Kühnle and Oberfichtner, 2017; Jessen et al., 2019; Schober and Spieß, 2013; Schober and Stahl, 2014; Scholz et al., 2019). This, in turn, implies that the alternative care environment at home is likely of higher quality for these children than for children from families with higher resistance to center-based child care (see e.g. Knudsen et al., 2006). In all likelihood, this mechanism leads to smaller treatment effects for children from low-resistance families.

Still, from a policy perspective, the main message conveyed by Chapter 2 is that center-based child care is not only not detrimental to child development but can also have long-time positive effects. Since the positive effects mostly materialize among weaker students, especially policy-makers concerned with fighting inequality should find a publicly financed expansion of child care an attractive policy option. However, while these general policy implications are undisputed, a closer look at the situation in Germany raises some doubts about their practical value. The main point is that today virtually all 3-year-olds attend child care. In other words, the obvious policy implication has already been put in place. In fact, the very reform that is studied in Chapter 2 has probably contributed most to this state of affairs. A logical follow-up question would then relate to the validity of the results for even younger children whose child care participation is still more mixed. It is reassuring that – at least in the short-run – positive effects of child care attendance between 0 and 2 have also been found in Germany. What is more, these effects disproportionately accrue to disadvantaged children as is the case in the present study (see Felfe and Lalive, 2013; 2018). To the best of my knowledge, credible long-run studies from Germany do not exist, yet.

Chapter 3 investigates if achievement in primary school can be increased by more intensive use of engaging teaching practices. While no effect is found in the full sample, subgroup-specific results suggest that children from low socio-economic backgrounds are positively affected. Subject-specific analyses reveal that engaging teaching practices are especially beneficial in reading.

When interpreting the results, one must pay particular attention to the makeup of the ESL scale, the main treatment indicator. Two issues have to be discussed in this context. First, there is no natural counterpart to engaging teaching practices as, for instance, is the case in studies on ‘modern’ student-centered teaching practices versus ‘traditional’ teacher-centered methods. Since most engaging teaching practices revolve around rather time-consuming activities such as repetition, summarizing, questioning, encouraging, and praising, it is argued that such practices most likely crowd out additional content. However, other alternative activities are also conceivable, for instance writing more tests and exams. Against this background, the detected effect has to be considered as an effect of engaging teaching practices against all other teaching methods that are applied in German primary schools. Second, the data do unfortunately not contain subject-specific information on teachers' use of engaging teaching practices. Instead, there is only one subject-independent indicator value per teacher. This introduces measurement error if there is within-teacher variation in the use of engaging teaching practices. The resulting attenuation bias depresses the estimated coefficients towards zero, which means that the estimated effects may constitute lower bounds on the true effects.

From a methodological standpoint, subject-specific values on the ESL scale would have allowed me to estimate within-teacher models that rule out bias due to unobserved personality differences between teachers. This would have been favorable, since despite the fact that the data provide rich socio-demographic and teaching style-related background information on instructors, omitted variable bias due to unobserved personality differences cannot entirely be ruled out. This concern is related to the question what in fact *causes* the existing variation in the use of engaging teaching practices by different teachers if not differences in their personalities. One possibility is different foci in the training that teachers have undergone, which may result from differences in the personalities of teacher-trainers or, more generally, from general differences in teacher training at different times and in different places. However, since teacher personalities can neither be captured entirely via control variables nor can they be eliminated methodologically, the results of Chapter 3 cannot strictly be interpreted as causal evidence on the use of engaging teaching practices.

To the extent that the detected effects are not simply the result of personality differences between teachers, employing engaging teaching practices can be used to improve learning outcomes of students from low socio-economic backgrounds. Crucially, policymakers are not faced with a direct trade-off between favoring children from high or low socio-economic backgrounds since the latter are not significantly negatively affected by engaging teaching practices. At this point, it is important to recall that the alternative to using engaging teaching practices, while not precisely defined, is what is actually happening in primary school classrooms around Germany. The estimated gains for children from low socio-economic backgrounds could therefore be realized rather simply and inexpensively by slightly altering the approach to teaching and – for future generations of teachers – the focus of teacher training. Since such interventions in teacher training cannot easily be priced, cost-benefit analyses cannot be conducted in the present setting. However, it should be noted that, in theory, changing the focus of teacher training should not be excessively expensive as compared to other policy interventions, especially those that influence personnel costs such as hiring additional teachers.

Chapter 4 deals with the relationship between class size and achievement in primary school. The results suggest that smaller classes are beneficial for learning outcomes in language and math and reduce the risk of repeating a grade. This is the first causal evidence on positive effects of class size reductions in Germany. The association between class size and achievement is driven by strong positive effects of class size reductions in large classes, i.e. classes with more than 20.5 students. As in chapters 2 and 3, certain groups of disadvantaged students benefit the most. Finally, Chapter 4 also makes a theoretical contribution to the literature by demonstrating how the initial size of a cohort is mechanically related to the student composition in higher grades. If not properly addressed, this depresses estimates based on within-school variation in cohort size toward zero.

When attempting to translate the findings of Chapter 4 into policy recommendations, a central limitation is the analysis' focus on short-run effects (a feature shared with the analysis in Chapter 3). With the data at hand we are not able to draw definite conclusions on whether the detected effects persist, diminish, vanish, or even increase over the medium- and long-run. Such conclusions, however, would be crucial to assess the impact of class size interventions on cognitive skill levels of the labor force, which is what economists are ultimately interested in. The task of speculating about possible long-run effects is hampered by the fact that we do not observe what mechanisms can explain our results. Possible mechanisms are for instance fewer disruptions in smaller classes or the possibility to apply more efficient teaching methods. Information about such mechanisms could provide a starting point of a discussion on whether class size effects persist and also on whether we can expect a similar relationship between class size and achievement in secondary school.

The above discussion notwithstanding, there are two reasons to believe that short-run gains from smaller classes *do* actually translate into long-run benefits. First, we know that cognitive skills are particularly malleable before the age of 10, after which at least general intelligence is relatively stable. Since this age threshold coincides with the end of primary schooling in most German states, we are confident that the detected effects last. Second, while long-run studies on class size are rare, existing evidence from Sweden suggests that class size interventions in primary school can have effects on wages and earnings that last well into adulthood (Fredriksson et al., 2013).

Nevertheless, even when conceding that positive long-run effects are likely, it is not easy to draw policy conclusions. This is due to the fact that class size reductions are costly, since additional teachers have to be hired and additional classrooms have to be provided. Ideally, one would therefore carry out analyses that weigh the additional costs against the expected benefits due to higher wages and earnings. In the absence of precise estimates on long-run effects, this is, however, impossible. Again, the only benchmark in terms of the efficiency of class size reductions comes from Sweden and is provided by Fredriksson et al. (2013). They estimate an internal rate of return of class size reductions of between 0.089 and 0.178. This is reassuring for the present analyses, since, in the same study, Fredriksson et al. (2013) also estimate effects on cognitive skills at the end of primary school that are only slightly larger than our estimates.

Against this background, the most modest policy implication would be a warning that reforms aimed at saving money via class size increases do not come for free, but have a cost in terms of significantly lower student achievement. These costs are particularly large in larger classes. Of course, the opposite is true for class size reductions, which cost money, but deliver benefits in the form of achievement gains. Since we find little evidence of class size effects in smaller classes, our results also suggest that any further class size reductions below a threshold of around 20.5 students per class have no effect. On the contrary, one has to conclude that class size may be increased up to a certain size without negative consequences for student achievement. A further policy implication pertains to the particularly large effects on disadvantaged children. This finding warrants the use of progressive maximum class size rules that prescribe smaller maximum class sizes as the number of disadvantaged children in a grade increases. Saarland is one of several German states that practices these flexible rules.

Next to the individual findings of each study in this dissertation, there are a number of ***common findings*** that warrant mention. *First*, all considered interventions are particularly beneficial for certain groups of disadvantaged students. Disadvantages for children may arise for very different reasons. In this dissertation, I have looked at students who are disadvantaged because their parents have low levels of education, students who are disadvantaged because of their poor command of the German language, students who are disadvantaged because of learning disabilities, and students who are disadvantaged because they simply perform worse than their peers for no apparent reason. Not all disadvantages lead to larger effect sizes on all considered interventions; but for all interventions the largest effects are found among one disadvantaged group of students.

What do we learn from this? At first glance, the results confirm earlier research, especially in the realm of early childhood education. On average, children from low socio-economic backgrounds stand to gain the most from attending child care because of their comparably worse alternative care quality at home (see e.g. Knudsen et al., 2006). However, the case is not that simple. For example, in Chapter 2 the largest effects are not found among students who are disadvantaged because of their socio-economic background, but among students who are disadvantaged because of their weak performance compared to other students. If differences in alternative care modes are still to explain this result, one would have to make the additional assumption that parents of lower achieving students provide worse care at home than other parents. Alternatively, it could be that the care mode in publicly-funded child care centers in Germany is more geared towards lower achieving children. However, in the absence of information on activities taking place in child care centers, this question cannot be investigated with the dataset used in Chapter 2.

Similar questions in regard to the interpretation of the results arise in respect to chapters 3 and 4. Here, the alternative to the treatment is not *no or less exposure* to formal education but rather a *different kind of exposure*. According to the technology of skill formation by Cunha and Heckman (2007), there are dynamic complementarities and self-productivity of skills. These dynamic processes imply that imparting new skills to students should be disproportionately effective among children from *high* socio-economic backgrounds, since they most likely dispose over higher baseline skill levels. The opposite is happening in the two present studies. One way to reconcile my results with the theory laid out by Cunha and Heckman (2007) is to conclude that the interventions examined in chapters 3 and 4 not only impart *new* skills to children but also lead to the transmission of skills that are only new to some students. For the rest of students – most likely the more advantaged students – complementarities between these latter skills and their existing skills are non-existent because of overlap. Taking a closer look at the interventions in question, this reading of the results seems plausible. In Chapter 3, I argue that a key characteristic of potentially engaging teaching practices is their focus on repetition and summarization which may crowd out additional exercises. This naturally favors students who do not master all exercises at once, i.e. weaker students. In Chapter 4, the largest effects are found among students with insufficient German proficiency or learning disabilities. Apparently, the special needs of these groups of students are better addressed in smaller classes. One way of reading this observation is that teachers devote more time to individual students and their (remedial) needs in smaller classes. Under the credible assumption that better students have fewer remedial needs, one comes to the conclusion that part of the effect of smaller classes is irrelevant for this group of students.

A *second* common finding of all three chapters is that in respect to the investigated interventions cognitive skill returns are larger in German language than in mathematics. At first glance, this seems to be at odds with some of the previous literature from the US that has found larger effect sizes of educational interventions in math than in reading (see e.g. Abdulkadiroglu et al., 2011; Angrist et al., 2012). An often-cited explanation for this finding is that reading achievement is harder to improve in schools since language is mostly developed outside of classrooms (Fryer, 2017). A second explanation posits that the critical period for language development occurs particularly early in life and that deficits can hardly be remedied later on (see e.g. Knudsen et al., 2006). There is ample evidence for this second explanation. For instance, Knudsen et al. (2006) demonstrate that language is most readily acquired before about 7 years of age, while Fryer (2017) finds a negative relationship between age and reading treatment effects in a meta-analysis of randomized education experiments. Importantly, this second explanation may also help reconcile my finding of larger treatment effects in German language than math with the literature from the US, because all interventions that are investigated in this dissertation start before children are 7 years old: In Chapter 2, the treatment occurs when children are three years old, in Chapter 3 between 6 and 10 years, and in Chapter 4 between 6 and 9 years.[124]

Taken together, the results of this dissertation shed some light on potential pathways towards higher student achievement, especially among disadvantaged children. They underscore that arising inequalities in cognitive skills can effectively be tackled at very young ages by very different means. While I have not performed specific cost-benefit analyses, there is reason to believe that implementing the investigated measures may be worthwhile from this angle, too. In Chapter 2, the estimated effects on children come on top of the well-known effects on maternal labor supply (for maternal employment effects of the same reform, see Bauernschuster and Schlotter, 2015), in Chapter 3 the costs of the intervention do not seem to be very high, and in Chapter 4 cost-benefit analyses from other countries provide encouraging benchmarks.

A task for future research will be to obtain more precise estimates of the efficiency of the proposed interventions, which then could form the basis for comparisons with alternative educational interventions that have not been discussed in this dissertation. However, such analyses critically hinge on the availability of better data. What is needed are datasets

[124] The duration of the treatment of several years in Chapters 3 and 4 is explained by the fact that most children experience the same teacher and class size throughout their entire time in primary school. The treatment therefore begins a lot earlier than when outcomes are measured, i.e. in grade 4 (Chapter 3) and grade 3 (Chapter 4), respectively.

that track individuals for an even longer time well into adulthood so that, eventually, information on lifetime wages can be linked to interventions in early childhood. Ideally, the interventions in these datasets would be based on randomized controlled field experiments such as the High Scope Perry Preschool Program or Project STAR in the US that have been touched upon in this dissertation. However, in the absence of well-implemented experiments on all possible interventions in education, a second-best data source are large household panel studies such as the SOEP or the NEPS. As these panels become longer over the coming years and decades, they will provide more complete information on the whole lifetimes of participating individuals that can be exploited by quasi-experimental methods.

BIBLIOGRAPHY

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and Student Achievement in Chicago Public High Schools. *Journal of Labour Economics, 25*, 95-135.

Abdulkadiroglu, A., Angrist, J. D., Dynarski, S. M., Kane, T. J., and Pathak, P. A. (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots. *Quarterly Journal of Economics, 126,* 699-748.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy, 113*, 151-184.

Ammermüller, A. and Pischke, J.-S. (2009). Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics*, *27*, 315–348.

Anderson, M. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*, 1481–1495.

Angrist, J. D., Battistin, E., and Vuri, D. (2017a). In a Small Moment: Class Size and Moral Hazard in the Italian Mezzogiorno. *American Economic Journal: Applied Economics, 9,* 216–249.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., and Walters, C. R. (2012). Who Benefits from KIPP? *Journal of Policy Analysis and Management*, *31*, 837-860.

Angrist, J. D. and Lavy, V. (1999). Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics, 114*, 533–575.

Angrist, J. D., Lavy, V., Leder-Luis, J., and Shany, A. (2017b). Maimonides Rule Redux. *NBER Working Papers 23486*, National Bureau of Economic Research, Inc.

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

Apps, P., Mendolia, S., and Walker, I. (2013). The impact of pre-school on adolescents' outcomes: Evidence from a recent English cohort. *Economics of Education Review, 37*, 183-199.

Argaw, B. A. and Puhani, P. A. (2018). Does class size matter for school tracking outcomes after elementary school? Quasi-experimental evidence using administrative panel data from Germany. *Economics of Education Review, 65*, 48–57.

Asadullah, M. N. (2005). The effect of class size on student achievement: evidence from Bangladesh. *Applied Economics Letters, 12*, 217–221.

Ashenfelter, O. and Zimmerman, D. J. (1997). Estimates of the Returns to Schooling from Sibling Data: Fathers, Sons, and Brothers. *The Review of Economics and Statistics, 79*, 1-9.

Aslam, M. and Kingdon, G. (2011). What Can Teachers Do to Raise Pupil Achievement? *Economics of Education Review, 30*, 559-574.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., and Blossfeld, H.-P. (2011). 4 Sampling designs of the National Educational Panel Study: challenges and solutions. *Zeitschrift für Erziehungswissenschaft, 14*, 51-65.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Econometrica, 74*, 431-497.

Autorengruppe Bildungsberichterstattung (2016). Bildung in Deutschland 2016. Technical report.

Bach, M., Koebe, J., and Peter, F. (2019). Long Run Effects of Universal Childcare on Personality Traits. mimeo.

Baker, M., Gruber, J., and Milligan, K. (2008). Universal Child Care, Maternal Labor Supply, and Family Well-Being. *Journal of Political Economy, 116*, 709-745.

Baker, M. (2011). Innis Lecture: Universal early childhoood interventions: what is the evidence base? *Canadian Journal of Economics, 44*, 1069-1105.

Barnett, W. S., Young, J. W., and Schweinhart, L. J. (1998). How preschool education influences long-term cognitive development and school success. In W.S. Barnett and S.S. Boocock (Eds.), *Early care and education for children in poverty: Promises, programs and long-term results*. 167-184. New York: State University of New York Press.

Barnett, W. S. (2011). Effectiveness of Early Educational Intervention. *Science, 333*, 975-978.

Baron, J. D. and Cobb-Clark, D. (2010). Are Young People's Educational Outcomes Linked to their Sense of Control? *IZA DP No. 4907.*

Bastos, P., Bottan, N. L., and Cristia, J. (2017). Access to Preprimary Education and Progression in Primary School: Evidence from Rural Guatemala. *Economic Development and Cultural Change, 65*, 521-547.

Bauchmüller, R., Gørtz, M., and Rasmussen, A. W. (2014). Long-Run Benefits from Universal High-Quality Pre-Schooling. *Early Childhood Research Quarterly, 29*, 457–470.

Bauernschuster, S. and Schlotter, M. (2015). Public child care and mothers' labor supply – Evidence from two quasi-experiments. *Journal of Public Economics, 123*, 1-16.

Becker, G. S. (1962). Investment in Human Capital: A Theoretical Analysis. *Journal of Political Economy, 70*, 9-49.

Berger, E., Peter, F. H., and Spieß, C. K. (2010). Wie hängen familiäre Veränderungen und das mütterliche Wohlbefinden mit der frühkindlichen Entwicklung zusammen? *Vierteljahreshefte zur Wirtschaftsforschung, 79*, 27-44.

Bietenbeck, J. (2014). Teaching Practices and Cognitive Skills. *Labour Economics, 30,* 143-153.

Bietenbeck, J., Ericsson, S., and Wamalva, F. (2017). Preschool attendance, school progression, and cognitive skills in East Africa. *IZA Discussion Paper No. 11212*.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2011). Too Young to Leave the Nest? The Effects of School Starting Age. *The Review of Economics and Statistics, 93*, 455–467.

Blanden, J., Del Bono, E., McNally, S., and Rabe, B. (2016). Universal pre-school education: The case of public funding with private provision. *Economic Journal, 126*, 682-723.

Blazar, D. (2015). Effective Teaching in Elementary Mathematics: Identifying Classroom Practices that Support Student Achievement. *Economics of Education Review, 48*, 16-29.

Borghans, L., Duckworth, A. L., Heckman, J. J., and ter Weel, B. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources, 43*, 972-1059.

Bos, W., Wendt, H., Koeller, O., and Selter, C. (2012a). *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im i nternationalen Vergleich*. Muenster/New York/Munich/Berlin: Waxmann.

Bos, W., Tarelli, I., Bremerich-Vos, A., and Schwippert, K. (2012b). *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Muenster/New York/Munich/Berlin: Waxmann.

Bowles, S., Gintis, H., and Osborne, M. (2001). The Determinants of Earnings: A Behavioral Approach. *Journal of Economic Literature, 91*, 1137-1176.

Bressoux, P., Kramarz, F., and Prost, C. (2009). Teachers' Training, Class Size and Students' Outcomes: Learning from Administrative Forecasting Mistakes. *Economic Journal, 119*, 540-561.

Bronars, S. G. and Oettinger, G. S. (2006). Estimates of the return to schooling and ability: Evidence from sibling data. *Labour Economics, 13*, 19-34.

Browning, M. and Heinesen, E. (2007). Class Size, Teacher Hours and Educational Attainment. *Scandinavian Journal of Economics, 109*, 415–438.

Burgess, S.M. (2016). Human Capital and Education: The State of the Art in the Economics of Education. *IZA DP. No. 9885*.

Camehl, G. (2018). *Non-cognitive Skills and the Quality of Early Education – Four Essays in Applied Microeconomics*. Inaugural Dissertation. Berlin: Free University.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., and Miller-Johnson, S. (2002). Early Childhood Education: Young Adult Outcomes From the Abecedarian Project. *Applied Developmental Science, 6,* 42-57.

Carlsson, M., Dahl, G. B., Öckert, B., Rooth, D.-O. (2015). The Effect of Schooling on Cognitive Skills. *The Review of Economics and Statistics, 97*, 533-547.

Carneiro, P. and Ginja, R. (2014). Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start. *American Economic Journal: Economic Policy, 6*, 135-173.

Carneiro, P. and Heckmann, J. J. (2003). Human capital policy. In: J. J. Heckman, A. B. Krueger, and B. M. Friedman (Eds.), *Inequality in America: What role for human capital policies?* pp. 77-240. Cambridge, MA: MIT Press.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Cascio, E. U. (2009). Do investments in universal early education pay off? Long-term effects of introducing kindergartens into public schools. *NBER Working Paper No. 14951*.

Cascio, E. U. and Schanzenbach, D. W. (2013). The impacts of expanding access to high-quality preschool education. *Brookings Papers on Economic Activity*, 127–178.

Cawley, J., Heckman, J. J., and Vytlacil, E. J. (2001). Three observations on wages and measured cognitive ability. *Labour Economics, 8*, 419-442.

Chamberlain, G. (1982). Multivariate Regression Models for Panel Data. *Journal of Econometrics, 18,* 5-46.

Chamberlain, G. (1984). Panel Data. In: Z. Griliches and M.D. Intriligator (Eds.), *Handbook of Econometrics. Volume 2*. 1247-1318. Amsterdam: North Holland.

Chambers, B., Cheung, A., Slavin, R. E., Smith, D., and Laurenzano, M. (2010). Effective Early Childhood Education Programs: A Systematic Review. Center for Data-Driven Reform in Education. John Hopkins University School of Education.

Chartier, A. M. and Geneix, N. (2007). Pedagogical approaches to early childhood education. Background Paper. UNESCO.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics, 126*, 1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review, 104,* 2633-2679.

Cho, H., Glewwe, P., and Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review, 31*, 77–95.

Ciccone, A. and Garcia-Fontes, W. (2015). Gender peer effects in school, a birth cohort approach. Economics Working Papers 1424, Department of Economics and Business, Universitat Pompeu Fabra.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *The Journal of Human Resources*, *41*, 778-820.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2010). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *Journal of Human Resources, 45*, 655-681.

Cobb-Clark, D. and Tan, M. (2011). Noncognitive skills, occupational attainment, and relative wages. *Labour Economics, 18*, 1-13.

Cohen-Zada, D., Gradstein, M., and Reuven, E. (2013). Allocation of students in public schools: Theory and new evidence. *Economics of Education Review, 34*, 96–106.

Crane, J. and Barg, M. (2003). Do Early Childhood Intervention Programs Really Work? Coalition for Evidence-Based Policy.

Cunha, F. and Heckman, J. J. (2007). The Technology of Skill Formation. *American Economic Review, 97*, 31-47.

Cunha, F., Heckman, J. J., Lochner, L. J., and Masterov, D. V. (2006). Interpreting the Evidence on Lifecycle Skill Formation. In E.A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, 697-812. Amsterdam: North Holland.

Currie, J. (2001). Early Childhood Education Programs. *Journal of Economic Perspectives, 15*, 213–238.

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

Cornelissen, T., Dustmann, C., Raute, A., and Schönberg, U. (2018). Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy, 126*, 2356-2409.

Cunha, F., Heckman, J., Lochner, L., and Masterov, D. (2006). Interpreting the Evidence on Life Cycle Skill Formation. In: E. Hanushek and F. Welch (Eds.). *Handbook of the Economics of Education*. 697-812. Amsterdam: North Holland.

Dahmann, S. C. (2016). *Human Capital Returns to Education. Three Essays on the Causal Effects of Schooling on Skills and Health*. Inaugural Dissertation. Free University of Berlin.

Danzer, N., Halla, M., Schneeweis, N., and Zweimüller, M. (2017). Parental Leave, (In)formal Childcare and Long-term Child Outcomes. IZA Discussion Paper No. 10812.

Datta Gupta, N. and Simonsen, M. (2010). Non-cognitive child outcomes and universal high quality care. *Journal of Public Economics, 94*, 30-43.

Datta Gupta, N. and Simonsen, M. (2016). Academic performance and type of early childhood care. *Economics of Education Review, 53*, 217-229.

Deary, I. J., Strand, S., Smith, P., and Fernandes, C. (2007). Intelligence and Educational Achievement. *Intelligence, 35*, 13-21.

DeCicca, P. and Smith, J. D. (2013). The long-run impacts of early childhood education: Evidence from a failed policy experiment. *Economics of Education Review, 36*, 41-59.

Dee, T. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review, 95,* 158-165.

Denny, K. and Oppedisano, V. (2013). The surprising effect of larger class sizes: Evidence using two identification strategies. *Labour Economics, 23*, 57–65.

Dietrichson, J., Kristiansen, I. L., and Nielsen, B. C. V. (2018). Universal Preschool Programs and Long-Term Child Outcomes – A Systematic Review. VIVE – Viden til Velfaerd. Det Nationale Forsknings- og Analysecenter for Velfaerd. Kobenhavn.

Dobbelsteen, S., Levin, J., and Oosterbeek, H. (2002). The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition. *Oxford Bulletin of Economics and Statistics*, *64*, 17– 38.

Dohmen, T. and van Landeghem, B. (2019). Numeracy and Unemployment Duration. *IZA DP No. 12531*.

Drange, N. and Havnes, T. (2014). Early and bright? Child care for toddlers and early cognitive skills. In: Technical Report. mimeo.

Drange, N., Havnes, T., and Sandsor, A. M. J. (2016). Kindergarten for all: Long run effects of a universal intervention. *Economics of Education Review, 53,* 164-181.

Dumas, C. and Lefranc, A. (2012). Early schooling and later outcomes: Evidence from pre-school extension in France. In: Ermisch, J., M. Jäntti, and T. Smeeding (eds). *From Parents to Children: The Intergenerational Transmission of Advantage* (p. 164–189). New York: Russell Sage Foundation, 2012.

Ernst, A. (2017). Private communication.

European Commission (2011). *Grade retention during compulsory education in Europe: Regulations and statistics*. Education, Audiovisual and Culture Agency, European Commission.

Eurydice (2006). Das Bildungswesen in der Bundesrepublik Deutschland 2004. Technical report, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Fabricant, S. (1954). *Economic Progress and Economic Change*. 34th Annual Report. New York: National Bureau of Economic Research.

Falch, T., Sandsør, A. M. J., and Strøm, B. (2017). Do Smaller Classes Always Improve Students' Long-run Outcomes? *Oxford Bulletin of Economics and Statistics, 79,* 654–688.

Falck, O., Mang, C., and Wößmann, L. (2015). Virtually No Effect? Different Uses of Classroom Computers and their Effect on Student Achievement. *IZA DP No. 8939*.

Felfe, C. and Lalive, R. (2013). Early Child Care and Child Development: For Whom it Works and Why. SOEPpapers on Multidisciplinary Panel Data Research 536.

Felfe, C. and Lalive, R. (2018). Does early childcare affect children's development? *Journal of Public Economics*, *159*, 33-53.

Felfe, C., Nollenberger, N., and Rodríguez-Planas, N. (2015). Can't buy mommy's love? Universal childcare and children's long term cognitive development. *Journal of Population Economics*, *28,* 393–422.

Fessler, P. and Schneebaum, A. (2016). The returns to preschool attendance. Department of Economics Working Paper No. 233, Vienna University of Economics and Business.

Finn, J. D., Gerber, S. B., and Boyd-Zaharias, J. (2005). Small Classes in the Early Grades, Academic Achievement, and Graduating From High School. *Journal of Educational Psychology, 97,* 214–223.

Fort, M., Ichino, A., Zanella, G. (2017). The cognitive cost of daycare 0-2 for children in advantaged families. In: Technical report. mimeo.

Foy, P. (2013). TIMSS and PIRLS 2011 User Guide for the Fourth Grade Combined International Database. Boston: TIMSS & PIRLS International Study Center.

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research, 74,* 59-109.

Fredrikssson, P., Öckert, B., and Oosterbeek, H. (2013). Long-Term Effects of Class Size. *Quarterly Journal of Economics, 128*, 249-285.

Frick, J. R. and Goebel, J. (2011). Biography and Life History Data in the German Socio Economic Panel (SOEP, v27, 1984-2010). Data Documentation 61. Deutsches Institut für Wirtschaftsforschung. Berlin.

Fryer, R. G. Jr. (2017). The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. In E. Duflo and A. Banerjee (Eds.), *Handbook of Field Experiments. Vol. 2*. 95-322. Amsterdam: North-Holland.

Garces, E., Thomas, C., and Currie, J. (2002). Longer-Term Effects of Head Start. *American Economic Review, 92*, 999-1012.

García, J. L., Heckman, J. J., and Ziff, A. L. (2017). Gender Differences in the Benefits of an Influential Early Childhood Program. *IZA Discussion Paper No. 10758*.

Gary-Bobo, R. J. and Mahjoub, M.-B. (2013). Estimation of Class-Size Effects, Using Maimonides' Rule and Other Instruments: the Case of French Junior High Schools. Annals of Economics and Statistics, (111/112), 193–225.

Gilraine, M. (2018). Identifying Multiple Treatment from a Single Discontinuity: An Application to Class Size Caps.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., and Schupp, J. (2018). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics 2018*. Oldenbourg: De Gruyter.

Goldin, C. (2016). Human capital. In C. Diebolt and M. Haupert (Eds.), *Handbook of Cliometrics*, chapter 3, pp. 55-86. Berlin, Heidelberg: Springer-Verlag.

Goldhaber, D. D. and Brewer, D. J. (1997). Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. *The Journal of Human Resources, 32*, 505-523.

Goldhaber, D. D. and Anthony, E. A. (2007). Can Teacher Quality be Effectively Assessed? *The Review of Economics and Statistics, 89*, 134-150.

Goldsmith, A. H., Veum, J. R., and Darity, W. (1997). The impact of psychological and human capital on wages. *Economic Inquiry, 35*, 815-829.

Goodman, A. and Sianesi, B. (2005). Early education and children’s outcomes: How long do the impacts last? *Fiscal Studies, 26,* 513–548.

Grabner, R. H. and Stern, E. (2011). Measuring cognitive ability. In German Data Forum (RatSWD) (Ed.), *Building on progress. Expanding the research infrastructure for the social, economic, and behavioral sciences , Volume 2*. 753-768. Opladen: Budrich UniPress Ltd.

Green, D. A. and Riddell, W. C. (2003). Literacy and earnings: An investigation of the interaction of cognitive and unobserved skills in earnings generation. *Labour Economics, 10*, 165-184.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature, 24*, 1141–1177.

Hanushek, E. A. (1989). Expenditures, Efficiency, and Equity in Education: The Federal Government’s Role. *The American Economic Review, 79,* 46–51.

Hanushek, E. A. (1996). A More Complete Picture of School Resource Policies. *Review of Educational Research, 66,* 397–409.

Hanushek, E. A. (1998). The Evidence on Class Size. Wallis Working Papers WP10, University of Rochester – Wallis Institute of Political Economy.

Hanushek, E. A. (2011). The Economic Value of Higher Teacher Quality. *Economics of Education Review, 30*, 466-479.

Hanushek, E. A. and Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review, 100*, 267-271.

Hanushek, E. A. and Wößmann, L. (2008). The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature, 46*, 607-668.

Havnes, T. and Mogstad, M. (2011). No child left behind: Subsidized child care and children’s longrun outcomes. *American Economic Journal: Economic Policy, 3,* 97-129.

Havnes, T. and Mogstad, M. (2015). Is universal child leveling the playing field? *Journal of Public Economics, 127*, 100-114.

Hattie, J. (2009). *Visible Learning. A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. New York: Routledge.

Head Start (2013). Head Start Factsheet. Fiscal Year 2013. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/mr/factsheets/docs/hs-program-fact-sheet-2013.pdf.

Heckman. J. and Masterov, D. (2007). The productivity argument for investing in young children. *Science, 29*, 446-493.

Heckman, J. J. and Rubinstein, Y. (2001). The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review, 91*, 145-149.

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review, 103,* 2052-2086.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics, 24*, 411-480.

Heineck, G. and Anger, S. (2010). The returns to cognitive abilities and personality traits in Germany. *Labor Economics, 17*, 535–546.

Heineck, G. and Riphahn, R. T. (2009). Intergenerational Transmission of Educational Attainment in Germany – The Last Five Decades. *Journal of Economics and Statistics, 229*, 36-60.

Herbst, C. M. (2013). The impact of non-parental child care on child development: evidence from the summer participation dip. *Journal of Public Economics, 105*, 86-105.

Herbst, C. M. (2017). Universal child care, maternal employment, and children’s long-run outcomes: Evidence from the US Lanham Act of 1940. *Journal of Labor Economics, 35*, 519-564.

Hernstein, R. J. and Murray, C. A. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Hidalgo-Cabrillana, A. and López-Mayan, C. (2015). Teaching Styles and Achievement: Student and Teacher Perspectives. Working Paper 2/2015. Economic Analysis Working Paper Series. Universidad Autónoma de Madrid.

Hill, L. G. and Werner, N. E. (2006). Affiliative Motivation, School Attachment and Aggression in School. *Psychology in the Schools, 43*, 231-246.

Hopkins, K. D. and Bracht, G. H. (1975). Ten-Year Stability of Verbal and Nonverbal IQ Scores. *American Educational Research Journal, 12,* 469-477.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics, 115,* 1239–1285.

IEA (2011a). TIMSS 2011. Teacher Questionnaire. http://timssandpirls.bc.edu/timss2011/international-contextual-q.html.

IEA (2011b). PIRLS 2011. Teacher Questionnaire. Retrieved August 12, 2016 from http://timssandpirls.bc.edu/pirls2011/international-contextual-q.html.

Ikeda, M. and Garcia, E. (2014). Grade repetition: A comparative study of academic and non-academic consequences. OECD Journal: Economic Studies, 2013(1).

Jakubowski, M. and Sakowski, P. (2006). Quasi-experimental estimates of class size effect in primary schools in Poland. *International Journal of Educational Research, 45,* 202–215.

Jepsen, C. and Rivkin, S. (2009). Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size. *Journal of Human Resources, 44*, 223-250.

Jessen, J., Schmitz, S., and Waights, S. (2019). Understanding Day Care Enrolment Gaps. *DIW Discussion Papers, 1808*.

John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm Shift to the Integrative Big Five Trait Taxonomy. In O.P. John, R.W. Robins, and L.A. Pervin (Eds.), *Handbook of personality: Theory and research (3rd ed.).* 114-158. New York: Guilford Press.

Joncas, M. and Foy, P. (2013). Sample design in TIMSS and PIRLS. In: M.O. Martin and I.V.S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Jonen, G. and Eckhardt, T. (2006). Das Bildungswesen in der Bundesrepublik Deutschland 2004. Technical report, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Justice, L. M., Petscher, Y., Schatschneider, C., and Mashburn, A. (2011). Peer Effects in Preschool Classrooms: Is Children's Language Growth Associated With Their Classmates' Skills? *Child Development, 82*, 1768-1777.

Judge, T. A., Higgins, C. A., Thoresen, C. J., and Barrick, M. R. (1999). The Big Five Personality Traits, General Mental Ability, and Career Success Across The Life Span. *Personnel Psychology, 52*, 621-652.

Kane, T. J., Taylor, E. S., Tyler, J. H., and Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources, 46*, 587-613.

Khattab, N. (2015). Students' aspirations, expectations and school achievement: what really matters? *British Educational Research Journal, 41*, 731-748.

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

Koebe, J. and Spieß, C. K. (2019). Die frühe Bildung und Betreuung in Deutschland: Familien- oder Bildungspolitik oder beides? *Gesellschaft, Wirtschaft, Politik, 68*, 97-108.

Koedel, C., Mihaly, K., and Rockoff, J. E. (2015). Value-Added Modeling: A review. *Economics of Education Review, 47*, 180-195.

Krassel, K. F. and Heinesen, E. (2014). Class-size effects in secondary school. *Education Economics, 22*, 412–426.

Kreyenfeld, M., Spieß, C. K., and Wagner, G. G. (2000). Kindertageseinrichtungen in Deutschland: ein neues Steuerungsmodell bei der Bereitstellung sozialer Dienstleistungen. *DIW Wochenbericht, 18/2000,* 269-275.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *The Quarterly Journal of Economics, 114*, 497–532.

Krueger, A. B. (2003). Economic Considerations and Class Size. *The Economic Journal, 113*, F34–F63.

Krueger, A. B. and Whitmore, D. M. (2001). The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR. *Economic Journal, 111*, 1–28.

Kühnle, D. and Oberfichtner, M. (2017). Does Early Child Care Attendance Influence Children's Cognitive and Non-Cognitive Skill Development? *IZA Discussion Paper No. 10661*.

Kuger, S., Marcus, J., and Spieß, C. K. (2019). Day care quality and changes in the home learning environment of children. *Education Economics, 27*, 265-286.

Kultusministerkonferenz (2007). Vorgaben für die Klassenbildung. Schuljahr 2007/2008. http://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Klassenbildung_2007_Schulorg_Vorg.pdf.

Lang, F., Weiss, D., Stocker, A., and von Rosenbladt, B. (2007). Assessing Cognitive Capacities in Computer-Assisted Survey Research: Two Ultra-Short Tests of Intellectual Ability in the German Socio-Economic Panel (SOEP). *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften, 127*, 183–192.

Lavy, V. (2015). What Makes an Effective Teacher? Quasi-Experimental Evidence. *CESifo Economic Studies, 62*, 88-125.

Lavy, V. (2015). Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries. *The Economic Journal*, *125,* F397-F424.

Leuven, E. and Oosterbeek, H. (2018). Class size and student outcomes in Europe. Technical report, European Expert Network on Economics of Educations.

Leuven, E., Oosterbeek, H., and Rønning, M. (2008). Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway. *Scandinavian Journal of Economics*, *110*, 663–693.

Lindqvist, E. and Vestman, R. (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics, 3*, 101-128.

Lorena Comi, S., Argentin, G., Gui, M., Origo, F., and Pagani, L. (2017). Is it the Way They Use it? Teachers, ICT and Student Achievement. *Economics of Education Review, 56*, 24-39.

Ludwig, J. and Miller, D. L. (2007). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly Journal of Economics, 122*, 159-208.

Lundberg, S. (2013). The College Type: Personality and Educational Inequality. *Journal of Labor Economics*, *31*, 421-441.

Martin, M. O., Mullis, I. V. S., Foy, P., and Stanco, G. M. (2012). *TIMSS 2011 International Results in Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O. and Mullis, I. V. S. (Eds.). (2013). TIMSS and PIRLS 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade – Implications for Early Learning. Boston: TIMSS & PIRLS International Study Center.

McCoy, S., Smyth, E., and Banks, J. (2012). The Primary Classroom: Insights from the Growing Up in Ireland Study. Learning in Focus. The Economic and Social Research Institute. Dublin.

McCrae, R. and Costa, P. J. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. Wiggins (Ed.), The Five Factor Model of Personality: Theoretical Perspectives. 51-87. New York: Guilford.

McLaughlin, M., McGrath, D. J., Burian-Fitzgerald, A., Lanahan, L., Scotchmer, M., Enyeart, C., and Salganik, L. (2005). Student Content Engagement as a Construct for the Measurement of Effective Classroom Instruction and Teacher Knowledge. American Institutes for Research. http://www.air.org/sites/default/files/downloads/report/AERA2005Student_Content_Engagement11_0.pdf.

Metzler, J. and Wößmann, L. (2012). The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation. *Journal of Development Economics, 99,* 486-496.

Mincer, J. (1958). Investments in Human Capital and Personal Income Distribution. *Journal of Political Economy, 66*, 281–302.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. Cambridge: NBER.

Mitchell, L., Wylie, C., and Carr, M. (2008). Outcomes of Early Childhood Education: Literature Review. New Zealand Council for Educational Research. Report to the Ministry of Education.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., and Ehrle, K. (1999). Evaluating the SAGE program: A pilot program in targeted pupil-teacher reduction in Wisconsin. *Educational Evaluation and Policy Analysis, 21*, 165–177.

Mueller, G. and Plug, E. (2006). Estimating the Effect of Personality on Male and Female Earnings. *Industrial and Labor Relations Review, 60,* 3-22.

Müller, K.-U., Spieß, C. K., Tsiasioti, C., Wrohlich, K., Bügelmayer, E., Haywood, L., Peter, F., Ringmann, M., and Witzke, S. (2013). Evaluationsmodul: Förderung und Wohlergehen von Kindern. DIW Berlin, Politikberatung kompakt 73.

Mulligan, C. B. and Sala-i Martin, X. (2000). Measuring Aggregate Human Capital. *Journal of Economic Growth, 5*, 215-252.

Mullis, I. V. S., Martin, M. O., Foy, P., and Arora, A. (2012a). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., and Drucker, K. T. (2012b). *PIRLS 2011 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica, 46*, 69-85.

Murnane, R. J. and Phillips, B. (1981). What do Effective Teachers of Inner-City Children Have in Common? *Social Science Research, 10*, 83-100.

Nandrup, A. B. (2016). Do class size effects differ across grades? *Education Economics, 24*, 83–95.

Neisser, U. C., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S.J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., and Urbina, S. (1996). Intelligence: Knowns and Unknowns. *American Psychologist, 51*, 77–101.

Noboa Hidalgo, G. and Urzua, S. (2012). The effect of participation in public childcare centers: evidence from Chile. *Journal of Human Capital, 6*, 1-34.

OECD (2004). Early Childhood Education and Care Policy in The Federal Republic of Germany. OECD Country Note. Technical Report.

OECD (2011). When Students Repeat Grades or are Transferred out of School: What Does it Mean for Education Systems? Technical report, PISA in Focus, educational policy brief.

OECD (2019). *Education at a Glance 2019: OECD Indicators*. Paris: OECD Publishing.

Osborne Groves, M. (2005). How important is your personality? Labor market returns to personality for women in the US and UK. *Journal of Economic Psychology, 26,* 827-841.

Paredes, V. (2014). A Teacher Like me or a Student Like me? Role Model versus Teacher Bias Effect. *Economics of Education Review, 39,* 38-49.

Paulus, C. and Leidinger, M. (2009). Landesweite Orientierungsarbeiten in der Grundschule im Saarland. Technical report, FR Erziehungswissenschaft der Universität des Saarlandes.

Perdue, N. H., Manzeske, D. P., and Estell, D. B. (2009). Early Predictors of School Engagement: Exploring the Role of Peer Relationships. *Psychology in the Schools, 46*, 1084-1097.

Piketty, T. and Valdenaire, M. (2006). L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français – Estimations à partir du panel primaire 1997 et du panel secondaire 1995. Post-Print halshs-00754847, HAL.

Piopiunik, M. and Schlotter, M. (2012). Identifying the Incidence of "Grading on a Curve": A Within-Student Across-Subject Approach. Ifo Working Paper No. 212. Munich.

Puma, M., Bell, S., Cook, R., and Heid, C. (2010). Head Start Impact Study. Final Report. Office of Planning, Research, and Evaluation, Administration of Children and Families. U.S. Department of Health and Human Services. Washington, DC.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80,* 1-28.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73,* 417-458.

Rockoff, J. (2009). Field Experiments in Class Size from the Early Twentieth Century. *Journal of Economic Perspectives, 23*, 211–230.

Ruhm, C. and Waldfogel, J. (2011). Long-Term Effects of Early Childhood Care and Education. IZA Discussion Paper No. 6149. Bonn.

Sandner, M. (2013). Effects of Early Childhood Intervention on Child Development and Early Skill Formation. Evidence from a Randomized Control Trial. Leibniz Universität Hannover. Faculty of Economics and Management. Discussion Paper 518.

Schacter, J. and Thum, Y. M. (2004). Paying for High- and Low-Quality Teaching. *Economics of Education Review, 23*, 411-430.

Schlotter, M. and Wößmann, L. (2010). Frühkindliche Bildung und spätere kognitive und nichtkognitive Fähigkeiten: Deutsche und internationale Evidenz. *DIW-Vierteljahreshefte zur Wirtschaftsforschung, 79*, 99-120.

Schlotter, M. (2011). The Effect of Preschool Attendance on Secondary School Track Choice in Germany. Evidence from Siblings. Ifo Working Paper No. 106.

Schneider, S. L. (2010). Nominal comparability is not enough: (In-)equivalence of construct validity of cross-national measures of educational attainment in the European Social Survey. *Research in Social Stratification and Mobility, 28*, 343-357.

Schober, P. S. and Spieß, C. K. (2013). Early Childhood Education Activities and Care Arrangements of Disadvantaged Children in Germany. *Child Indicators Research, 6,*

709-735.

Schober, P. S. and Stahl, J. F. (2014). Childcare Trends in Germany: Increasing Socio-Economic Disparities in East and West. *DIW Economic Bulletin, 11*, 51-58.

Scholz, A., Erhard, K., Hahn, S., and Harring, D. (2019). *Inequalities in Access to Early Childhood Education and Care in Germany. The Equal Access Study*. ICEC Working Paper Series – Volume 2. Munich: Deutsches Jugendinstitut e.V.

Schultz, T. W. (1961). Investment in Human Capital. *American Economic Review, 51*, 1–17.

Schwerdt, G. and Wuppermann, A. C. (2011). Is Traditional Teaching Really All That Bad? A Within-Student Between-Subject Approach. *Economics of Education Review, 30,* 365-379.

Seidel, T. and Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research, 77*, 454-499.

Shernoff, D. J. (2013). Optimal Learning Environments to Promote Student Engagement. New York: Springer.

Smith, A. (1776). An Inquiry into the Nature and Causes of the Wealth of Nations, Book II: Of the Nature, Accumulation, and Employment of Stock. https://en.wikisource.org/wiki/The_Wealth_of_Nations.

Smith, A. (2015). The long-run effects of universal pre-K on criminal activity. Unpublished manuscript. https://ssrn.com/abstract=2685507.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics, 87,* 355-374.

Solow, R. M. (1957). Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics, 39,* 312-320.

Spieß, C. K. (2008). Early Childhood Education and Care in Germany: The Status Quo and Reform Proposals. *Zeitschrift für Betriebswirtschaftslehre, 67,* 1-20.

Spieß, C. K. (2011). Vereinbarkeit von Familie und Beruf – wie wirksam sind deutsche "Care Policies"? *Perspektiven der Wirtschaftspolitik, 12,* 4-27.

Spieß, C. K. (2017). Early Childhood Education and Care Services and Child Development: Economic Perspectives for Universal Approaches. In: R.A. Scott, S.M. Kosslyn, N. Pinkerton (Eds.), *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*. Wiley Online Library.

Statistisches Amt des Saarlands (2017). Personal Provision of Data.

Statistisches Landesamt Sachsen (2017). Personal Provision of Data.

Statistisches Bundesamt (2010). *Allgemeinbildende Schulen. Schuljahr 2007/8*. Fachserie 11 Reihe 1. Wiesbaden.

Statistisches Bundesamt (2013). Kinder- und Jugendhilfestatistik: Betreuungsquoten der Kinder unter 6 Jahren in Kindertagesbetreuung am 01.03.2013. Wiesbaden: Statistisches Bundesamt.

Stevenson, P. R. (1922). Relation of Size of Class to School Efficiency. *University of Illinois Bulletin, 14*, 1–39.

Thiel, H. and Thomsen, S. (2013). Noncognitive skills in economics: Models, measurement, and empirical evidence. *Research in Economics, 67*, 189-214.

Tyler, J. H., Taylor, E. S., Kane, T. J., and Wooten, A. (2010). Using Student Performance Data to Identify Effective Classroom Practices. *American Economic Review, 100*, 256-260.

Urquiola, M. (2006). Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia. *The Review of Economics and Statistics, 88*, 171–177.

Urquiola, M. and Verhoogen, E. (2009). Class-Size Caps, Sorting, and the Regression-Discontinuity Design. *American Economic Review, 99*, 179–215.

van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

van Klaveren, C. (2011). Lecturing Style Teaching and Student Performance. *Economics of Education Review, 30*, 729-739.

Vogel, C. A., Xue, Y., Moiduddin, E. M., and Carlson, B. L. (2010). Early Head Start Children in Grade 5: Long-Term Follow-Up of the Early Head Start Research and Evaluation Project Study Sample. OPRE Report 2011-8. Office of Planning, Research, and Evaluation, Administration of Children and Families. U.S. Department of Health and Human Services. Washington, DC.

Wagner, G. G., Frick, J. R., and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP): Scope, evolution and enhancements. *Schmollers Jahrbuch, 127*, 139-169.

Wößmann, L. (2003a). Specifying Human Capital. *Journal of Economic Surveys, 17*. 239–270.

Wößmann, L. (2003b). Schooling Resources, Educational Institutions, and Student Performance: The International Evidence. *Oxford Bulletin of Economics and Statistics, 65*, 117-170.

Wößmann, L. (2005). Educational production in Europe. *Economic Policy, 20*, 445–504.

Wößmann, L. (2016). The Economic Case for Education. *Education Economics, 24*, 3-32.

Wößmann, L. and West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review, 50*, 695–736.

Yoshizawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Magnuson, K. A., Phillips, D., and Zaslow, M. J. (2013). Investing in

Our Future: The Evidence Base on Preschool Education. Research Brief. Society for Research in Child Development. Foundation for Child Development.