DIW BERLIN

1817

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2019

Birth Cohort Size Variation and the Estimation of Class Size Effects

Maximilian Bach and Stephan Sievert

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Birth Cohort Size Variation and the Estimation of Class Size Effects*

Maximilian Bach† Stephan Sievert‡

July 26 2019

Abstract

We present evidence that the practice of holding back poorly performing students affects estimates of the impact of class size on student outcomes based on within-school variation of cohort size over time. This type of variation is commonly used to identify class size effects. We build a theoretical model in which cohort size is subject to random shocks and students whose performance falls below a threshold are retained. Our model predicts that initial birth cohort size is mechanically related to the grade-level share of previously retained students once these cohorts reach higher grades. This compositional effect gives rise to an upward bias in class size effects exploiting variation in birth cohort size. Using administrative data on school enrollment for all primary schools in one federal state of Germany, we find support for this compositional effect. Correcting for the resulting bias in a unique dataset on standardized test scores for the full student population of third graders, we find that not only are smaller classes beneficial for language and math test scores, but also for reducing grade repetition.

Keywords: Class size effects; Quasi-experimental evidence; Student achievement; Primary school

JEL Codes: I20, I21, I29

*We benefited from helpful discussions with Edwin Leuven, Helena Holmlund, Jan Nimczik, Bernd Fitzenberger, Jan Marcus, Felix Weinhardt, C. Katharina Spieß, Mathias Hübener, Jochen Pröbsting, Sophia Schmitz, Julia Schmieder, Jan Berkes, as well as with the audiences at ZEW Berlin, DIW Berlin, Hertie School of Governance, HU Berlin, and the conferences BIEN and IWAEE. We also thank Friedhelm Münch und Christoph Paulus for providing a subset of the data.

†Corresponding author. DIW Berlin & Humboldt University Berlin. Email: mbach@diw.de

‡Free University Berlin

1 Introduction

Class size is one of the most important determinants of the costs of education as teachers' salaries account for the bulk of costs in public education in most countries. 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 of 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).[1]

This paper offers an 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.[2]

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

[1] 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).

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

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 — increases average test scores in these classes. Since grade retention or delayed enrollment in primary school is a common practice in most countries,[3] 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 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

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

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.[4] 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 sizes 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 of the true effect sizes. Our study provides the first causal evidence of significant class size effects on test scores in Germany.[5] The 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

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

[5]Previous quasi-experimental studies for Germany cannot conclude that smaller classes improve student achievement. Wößmann (2005) is the only study for Germany 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 classes.

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 disorder 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, if anything, 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 re-visit this empirical strategy to further investigate class size effects in other contexts. This 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 studies to investigate the effects of class size.[6] First, the within-school design is widely applicable and allows for studying class size effects even if no class size

[6]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.

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.[7] Second, regression discontinuity designs (RDD) can yield biased estimates in some contexts where carefully implemented within-school designs may not.[8] 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 class size effects we find in our data, which is missed in RDDs.

The rest of the paper is organized as follows: Section 2 discusses the related literature. Section 3 develops our theoretical model and its implications for previously used research designs. Section 4 sets out the institutional background for our empirical part. Section 5 presents our estimation strategy. Section 6 describes the data used in our analysis. Estimates are presented and interpreted in section 7, with conclusions drawn in section 8.

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.[9] The methods applied in these studies can be broadly classified into three categories. The first is randomized experiments. Tennessee's Student

[7] 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.

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

[9] 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.

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 more students. 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.[10] 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 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.[11] To isolate natural randomness in birth cohort sizes from any secular trends, she controls for very flexible school-district trends using 24 years of birth cohort data.[12] Her results indicate

[10]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.

[11]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.

[12]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

no class size effects and rule out effect sizes as small as 0.04 SD for a 10 percent reduction in class size.[13] 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.[14] We have listed all within-school studies that we could find and broken them down along a number of dimension in Table A.1. All studies use data from school systems that allow either for grade retention or redshirting of students.[15] 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.[16] In fact, of the 11 papers summarized, four find no significant class size effects and one even finds significant 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

school/expected-number-of-classes combination.

[13]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.

[14]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.

[15]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.

[16]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.

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.

3 Theoretical Model and Implications

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.[17] 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 test scores in higher grades.[18] 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

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

[18] 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.

to actual class size.[19] 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.

$$a_{is}^t > p_s^t \tag{1}$$

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.[20] We assume that the size and the grade retention threshold of cohorts are distributed with school-specific means

$$N_s^t = N_s + \eta_s^t \tag{2}$$

$$p_s^t = p_s + \nu_s^t \tag{3}$$

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 $Var(\nu_s^t) > 0$).[21] 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 $1/2\theta$ 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

$$\alpha_s^t = \alpha_s + \pi^\alpha N_s^t + \epsilon_s^t \tag{4}$$

where π^α is the effect of class size in LG on academic skills and ϵ_s^t are i.i.d. shocks

[19] 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 7.1.

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

[21] 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 **??**.

with mean zero and positive variance. In combination with the rule for grade retention in (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[22]

$$\lambda_s^t = \frac{\alpha_s^t + \theta - p_s^t}{2\theta} \tag{6}$$

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

$$N_{s\tau}^{obs} = \lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L-1}) N_s^{\tau-L-1} \tag{7}$$

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

$$\phi_s^\tau = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{N_{s\tau}^{obs}} = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{\lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L-1}) N_s^{\tau-L-1}} \tag{8}$$

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

$$\omega_{s\tau} = \tilde{\omega}_{s\tau} + \pi^\omega N_{s\tau}^{obs} \tag{9}$$

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 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

[22]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:

$$-\theta \leq \alpha_s^t - p_s^t \leq \theta \tag{5}$$

reach HG in year $\tau = t + L$ can be written as

$$E\left(test_{is}^{t}|non-retained\right) = E\left(test_{is}^{t}|a_{is}^{t} \geq p_{s}^{t}\right) = \frac{\alpha_{s}^{t} + \theta + p_{s}^{t}}{2} + \omega_{s,t+L} \tag{10}$$

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

$$\begin{aligned} E\left(test_{is}^{t}|retained\right) &= E\left(test_{is}^{t}|a_{is}^{t} < p_{s}^{t}\right) + \delta_{s}^{t} + \omega_{s,t+L+1} \\ &= \frac{\alpha_{s}^{t} - \theta + p_{s}^{t}}{2} + \delta_{s}^{t} + \omega_{s,t+L+1} \end{aligned} \tag{11}$$

where $E\left(a|a < p\right)$ denotes the average skills after L years in LG of students who were retained. The average test performance of all students in HG in year τ can be derived by combining (8), (10) and (11)

$$test_{s\tau} = \phi_{s}^{\tau-L} E\left(test_{is}^{\tau-L}|non-retained\right) + (1 - \phi_{s}^{\tau-L}) E\left(test_{is}^{\tau-L-1}|retained\right) \tag{12}$$

So far, we 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 childcare 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 timing of school enrollment, LG would refer to the last year in childcare 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.

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.[23] 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 the appendix, we show that if there are no class size effects this ratio is equal to

$$\frac{3(\theta - \delta)(1 - \lambda)\lambda}{3\lambda - 1} \tag{13}$$

where $(\theta - \delta)$ is the average test score difference of non-retained students and students retained in the past, see (10) and (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 the cohort t, N_s^t. We show that this positive shock translates into changes in the size of classes in HG as well as changes in the share of retained students in HG, which results in the spurious effect in (13). First note that this shock increases the number of

[23] 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 (2).

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 $t+L-1$.[24] 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 (13).[25] 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 $t+L+1$ and, therefore, further decrease the relative share of retained

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

[25] 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 (13) because it also appears in the numerator.

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 on 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.

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 $\{N_{s\tau}^{obs}, N_{s,\tau-1}^{obs}, test_{s\tau}, test_{s,\tau-1}, N_s^{\tau-L}, N_s^{\tau-L-1}\}$).[26]

The commonly used instrumental variable approach would estimate class size effects

[26]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.

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 $\tau - L$.[27] In the appendix, 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

$$\beta_{IV} = \underbrace{(\theta - \delta)\rho_{IV}}_{\text{grade retention bias I}} + \underbrace{\xi_{IV}}_{\text{attenuation factor}} \pi^{\alpha} + \pi^{\omega} \tag{14}$$

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.[28] 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.[29]

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.[30] The importance of this attenuation

[27]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.

[28]See the appendix for more details.

[29]Unlike expression (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 the appendix.

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

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.[31] In the appendix, we prove that by conditioning on whether a student has been retained the IV estimator will consistently estimate

$$\beta_{IV}^{REA} = \xi_{IV}\pi^{\alpha} + \pi^{\omega} \tag{15}$$

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 (15) thus yields a lower bound of the true class size effect.

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 size information to student test score data. Many studies in Table A.1, therefore, regress

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.

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

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 the appendix we show that in our set-up the resulting estimate will converge to

$$\hat{\beta}_{OLS} = \underbrace{(\theta - \delta)\rho_{OLS}}_{\text{grade retention bias I}} + \underbrace{\iota_{OLS}}_{\text{grade retention bias II}} + \underbrace{\xi_{OLS}}_{\text{attenuation factor}} \pi^{\alpha} + \pi^{\omega} \tag{16}$$

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 (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.[32] 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 counterparts, 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

$$\hat{\beta}_{OLS}^{REA} = \iota_{OLS} + \xi_{OLS}\pi^{\alpha} + \pi^{\omega} \tag{17}$$

However, ι_{OLS} does not disappear because it is a 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

[32]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 initital 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 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.[33] There is no explicit ability tracking in primary school.[34] 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 parentals' 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

[33]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).

[34]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 (Eurydice, 2006). 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 the 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 eventually merged, we make sure that the reform does not bias our estimates.

5 Estimation Strategy

The main difficulties in the identification of class size effects arise from students 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:

$$y_{icts} = \alpha_0 + \alpha_1 CS_{ts} + \alpha_2 X_i + T_t + S_s + \epsilon_{icts} \tag{18}$$

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 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 (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 the 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 (18).

As discussed in Section 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.[35] 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

[35]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.

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 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 (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. As our model predicts that the grade-level share of these students (who should have below average test scores) will be higher in years associated with larger initial birth cohorts. This should also upward bias our estimates.[36]

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 (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).[37] 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 third grade. 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

[36]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.

[37]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.

bias.[38]

Around one-third of all primary schools in Saarland were merged in 2005. This consolidation 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.[39]

As discussed in Section 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 (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 7.2 in which we test whether the composition of cohorts is systematically related to their size.

6 Data

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

[38]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.

[39]For efficiency reasons, we would ideally estimate only one set of fixed effects on 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 which school classes belonged to before consolidation. Hence, the need to aggregate everything to the consolidate school-level for merged schools.

state of Saarland.[40] [41] The data were obtained via state-wide centralized exams at the end of third grade in the school years 2003 to 2006. Participation in these "State-wide Orientation Exams" (SOE) was obligatory for all schools and classes.[42] Testing was carried out 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 the teacher 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 have been retained, who have been enrolled one year late and who have been enrolled one year early. This information is used to impute initial cohort size. Table

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

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

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

A.2 shows the structure of the Saarland data by academic year.

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 1 about here]

Table 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 independent 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 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 subsequent estimations.

[Figure 1 about here]

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 strongly correlates with a host of parental background measures such as income, education, and origin. In fact, Wößmann (2005) and Ammermueller and Pischke (2009) found 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 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 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 2 reports descriptive statistics on the fraction of students in the 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.

[Table 2 about here]

7 Results

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 3.

We use administrative enrollment data for grade 1 for Saarland and regress the fraction of 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 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.

[Table 3 about here]

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 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 A.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 A.4 show results for Saxony analogous to those reported in Table 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 3 by three.

The theoretical results in Section 3 imply that instrumental variable estimates will be biased if non-retained students have skills that differ, on average, from retained-, redshirted- 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 to 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.[43]

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

Thus, it allows identifying each group of students. Table 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.[44]

With the results from Tables 3 and 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 (15) we see that the bias is additive and equals the product of $(\theta - \delta)$ and ρ_{IV}.[45] 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 for retained students values of 0.564 ($= 3 \times 0.262 \times 0.717$) SD and 0.715 ($= 3 \times 0.262 \times 0.910$) SD for language and math, respectively.[46] 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.

[Table 4 about here]

[44] Another potential concern are students who skip a grade. Table 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.

[45] 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.

[46] The value 0.262 comes from column 3 of Panel B in Table 3. The second value, 0.717, is from row 3 and column 1 in Table 4. The second value for math comes from the second row of column 2 in Table 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.

7.2 Validity of birth cohort size variation

The key assumption of our estimation approach described in Section 5 is that within a school, 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 3 reports the results of these regressions. Reassuringly, the results indicate that early and late enrollment is balanced with respect to birth cohort size.[47] 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 correlations between the size and initial composition of birth cohorts.

[Table 5 about here]

In a second approach, we check for balancedness of student characteristics with respect to birth cohort size drawing on the student-level data. In Table 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 student skills in terms 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

[47] 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.

grade-level, the coefficients for being older and younger than typical third graders are significant and negative.[48] 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 enrolled 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.[49] 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.[50]

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 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

[48] 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.

[49] 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.

[50] As expected when running a number of regression 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.

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.

[Table 6 about here]

7.3 Class size effects

In this section, we turn to reporting our class size effects. Table 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-student-increase in predicted class size based on imputed cohort size leads to approximately a 0.45-student-increase in class size in grade 3.

[Table 7 about here]

Tables 8 contains our main results for the empirical model in (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.[51] The point estimates in both subjects are negative but not statistically significant. Our discussion of equation (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 (14) and (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.[52] The differences between estimates in columns 5 and 6 are not statistically significant and only the language

[51]The full regression results are reported in Tables A.5 -A.6 in the appendix.

[52]In Section 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 5, however, the differences in coefficients in column 5 and 6 are likely to understate any bias resulting from holding back poorly performing students. This is because

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.

[Table 8 about here]

Because students with insufficient German proficiency are, on average, placed in smaller classes in Saarland (see the discussion in Section 5 and Table 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 differences to the results. This suggest 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 without sufficient 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.

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 decreases class size estimates 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

we only condition on a proxy for whether or not a student has been held back in the past, which does not fully eliminate the bias resulting from these students. Therefore the implied size of the bias in Table 8 is a lower bound, explaining why it is slightly smaller than what we predicted.

retained students ensues, which offsets part of the negative correlation between class size and the share of held back students discussed before.[53] 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 (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 1 and 5 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 are 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 (15) and (17) imply that the estimates in Table 8 are attenuated because class size in grade 3 is not perfectly correlated with the class size students experienced in grades 1 and 2. [54]

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

[53]Unfortunately, comparing ρ_{IV} and ρ_{OLS} in equations (14) and (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 (D.19) in the appendix), which cannot be identified.

[54]Table A.7 reports estimates for different specification 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 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.

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 A.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 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 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.[55] In fact, any school with less than three years of data has to be dropped from the analysis. Hence we lose about 60 percent of observations.[56] The results of these regressions are reported in columns 2 and 5 of Table A.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

[55]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.

[56]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.

bias our class size effects positively, we expect our estimates without school-specific linear trends in Table 8 to provide lower bounds on the true class size effect.

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.

In Table 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.[57] 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

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

the same pattern of basically zero effects in small classes and large negative effects in larger classes for mathematics.[58]

[Table 9 about here]

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 to 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.[59]

7.3.2 Effect Heterogeneity

In our specifications in Tables 8 and 9 we implicitly assume that all students are similarly affected by class size. Krueger (1999), however, has shown 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

[58] As before, we also carry out robustness checks, such as including school-number of classes combination fixed effects, and school specific linear trends. Table A.10 in the Appendix reports results for the spline specification with a knot placed at 20.5. The results are qualitatively very similar, but as before, standard errors increase substantially.

[59] 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.

proficiency, reading disorder (dyslexia), and learning disability in math (dyscalculia). Table 10 shows the coefficients of these seven interactions.[60] In line with the hypothesis that disadvantaged students are harmed most by larger classes, all interaction terms, except for the female term, are negative and most are statistically significant at the 1 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 panel 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 our specifications in Tables 8 and 9 mask some marked effect heterogeneity for certain groups of students. Compared to the non-disadvantaged student, 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.

[Table 10 about here]

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 first grade for the 2001-2004 academic years.[61] We follow the same methodological approach as above, but now regress the

[60] Since the IV results are very similar we only report OLS results. For the IV results see Table A.11 in the appendix.

[61] 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

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 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 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 3, however, the estimate in column 3 could 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.[62] Since these students account for a larger share in larger cohorts within a shool, 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 estimated regressions where we used 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.[63] 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.

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.

[62]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.

[63]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.

Given that only 2.3 percent of all students repeat grade 1, this is an increase of almost 7 percent.[64] 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 implications for RDDs based on maximum class size rules. As retention rates increase with class size, marginal students with low 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 students 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.[65]

[Table 11 about here]

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

[64]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 1, which is the the population average for the 2001-2006 academic years.

[65]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 3 and assume that the class size effect on grade retention of 0.152 for grade 1 (from Table 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 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 type of composition bias laid out in Section 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.

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 once these cohorts progress through primary school. The reason is that the practice of holding back poorly performing students mechanically causes larger birth cohorts to be in grades with a smaller share of students who have been held back in the past. The resulting bias may help to 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 students has been held back a grade in the past — that produces a lower bound of the class size effect.

In the empirical part of this paper, we show that the two main empirical 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 effects 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 a cost in terms 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, this suggests that class size may be increased up to a certain size without negative consequences for student achievement. Second, larger benefits of smaller classes for disadvantaged children warrant the use of progressive maximum class size rules.

Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Kindergarten, doi:10.5157/NEPS:SC2:6.0.1. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

References

References

Ammermueller, 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(3):315–348.

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(4):216–249.

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(2):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.

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(4):217–221.

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(2):51.

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

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(2):455–467.

Bressoux, P., Kramarz, F., and Prost, C. (2009). Teachers' Training, Class Size and Students' Outcomes: Learning from Administrative Forecasting Mistakes. *Economic Journal*, 119(536):540–561.

Browning, M. and Heinesen, E. (2007). Class Size, Teacher Hours and Educational Attainment. *Scandinavian Journal of Economics*, 109(2):415–438.

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(4):1593–1660.

Cho, H., Glewwe, P., and Whitler, M. (2012). Do reductions in class size raise studentsÂ' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review*, 31(3):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.

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

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

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(1):17–38.

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,.

Falch, T., SandsÃ¸r, A. M. J., and StrÃ¸m, B. (2017). Do Smaller Classes Always Improve StudentsÂ' Long-run Outcomes? *Oxford Bulletin of Economics and Statistics*, 79(5):654–688.

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(2):214–223.

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 Discontinuty: An Application to Class Size Caps.

Hanushek, E. (1998). The Evidence on Class Size. Wallis Working Papers WP10, University of Rochester - Wallis Institute of Political Economy.

Hanushek, E. A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 24(3):1141–1177.

Hanushek, E. A. (1989). Expenditures, Efficiency, and Equity in Education: The Federal Government's Role. *The American Economic Review*, 79(2):46–51.

Hanushek, E. A. (1996). A More Complete Picture of School Resource Policies. *Review of Educational Research*, 66(3):397–409.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics*, 115(4):1239–1285.

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(3):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(1).

Jonen, G. and Eckhardt, T. (2006). Das Bildungswesen in der Bundesrepublik Deutschland 2004. Technical report, Sekretariat der StÃndigen Konferenz der Kultusminister der LÃnder in der Bundesrepublik Deutschland,.

Krassel, K. F. and Heinesen, E. (2014). Class-size effects in secondary school. *Education Economics*, 22(4):412–426.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *The Quarterly Journal of Economics*, 114(2):497–532.

Krueger, A. B. (2003). Economic Considerations and Class Size. *The Economic Journal*, 113(485):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(468):1–28.

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 fÃ̃¼r Wirtschafts- und Sozialwissenschaften*, 127(1):183–192.

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 RÃ¸nning, M. (2008). Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway. *Scandinavian Journal of Economics*, 110(4):663–693.

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.

Nandrup, A. B. (2016). Do class size effects differ across grades? *Education Economics*, 24(1):83–95.

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.

Paulus, C. and Leidinger, M. (2009). Landesweite Orientierungsarbeiten in der Grundschule im Saarland. Technical report, FR Erziehungswissenschaft der UniversitÃt des Saarlandes.

Piketty, T. and Valdenaire, M. (2006). L'impact de la taille des classes sur la rÃ©ussite scolaire dans les Ã©coles, collÃ¨ges et lycÃ©es franÃ§ais - Estimations Ã partir du panel primaire 1997 et du panel secondaire 1995. Post-Print halshs-00754847, HAL.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. (2009). Field Experiments in Class Size from the Early Twentieth Century. *Journal of Economic Perspectives*, 23(4):211–30.

Stevenson, P. R. (1922). Relation of Size of Class to School Efficiency. *University of Illinois Bulletin*, 14(45):1–39.

Urquiola, M. (2006). Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia. *The Review of Economics and Statistics*, 88(1):171–177.

Urquiola, M. and Verhoogen, E. (2009). Class-Size Caps, Sorting, and the Regression-Discontinuity Design. *American Economic Review*, 99(1):179–215.

van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

Wößmann, L. (2005). Educational production in Europe. *Economic Policy*, 20(43):445–504.

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(3):695–736.

Tables

Table 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.

Table 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. Source: Fachserie. 11, Bildung und Kultur. 1, Allgemeinbildende Schulen 2001/2002-2006/2007.

Table 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***	-0.164***	-0.045**
	(0.026)	(0.023)	(0.020)
	Panel B: IV grade composition		
Class size	-0.800***	-0.476***	-0.262***
	(0.081)	(0.073)	(0.055)
	Panel C: OLS birth cohort composition		
Imputed cohort size	0.029	0.002	
	(0.025)	(0.029)	
N SchoolYearObs	871	871	871

Notes: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

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

Table 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: 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

Table 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: Each column contains results for a separate regressions. Standard errors clustered at the combined school-level are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 7: First Stage Estimates

	Class size in grade 3			
	(1)	(2)	(3)	(4)
Class size predicted by imputed cohort size	0.446***	0.446***	0.446***	0.446***
	(0.034)	(0.034)	(0.034)	(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: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 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: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 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 ≥ 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 ≥ 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: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 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: 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 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, learnings disabilities, and native language. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

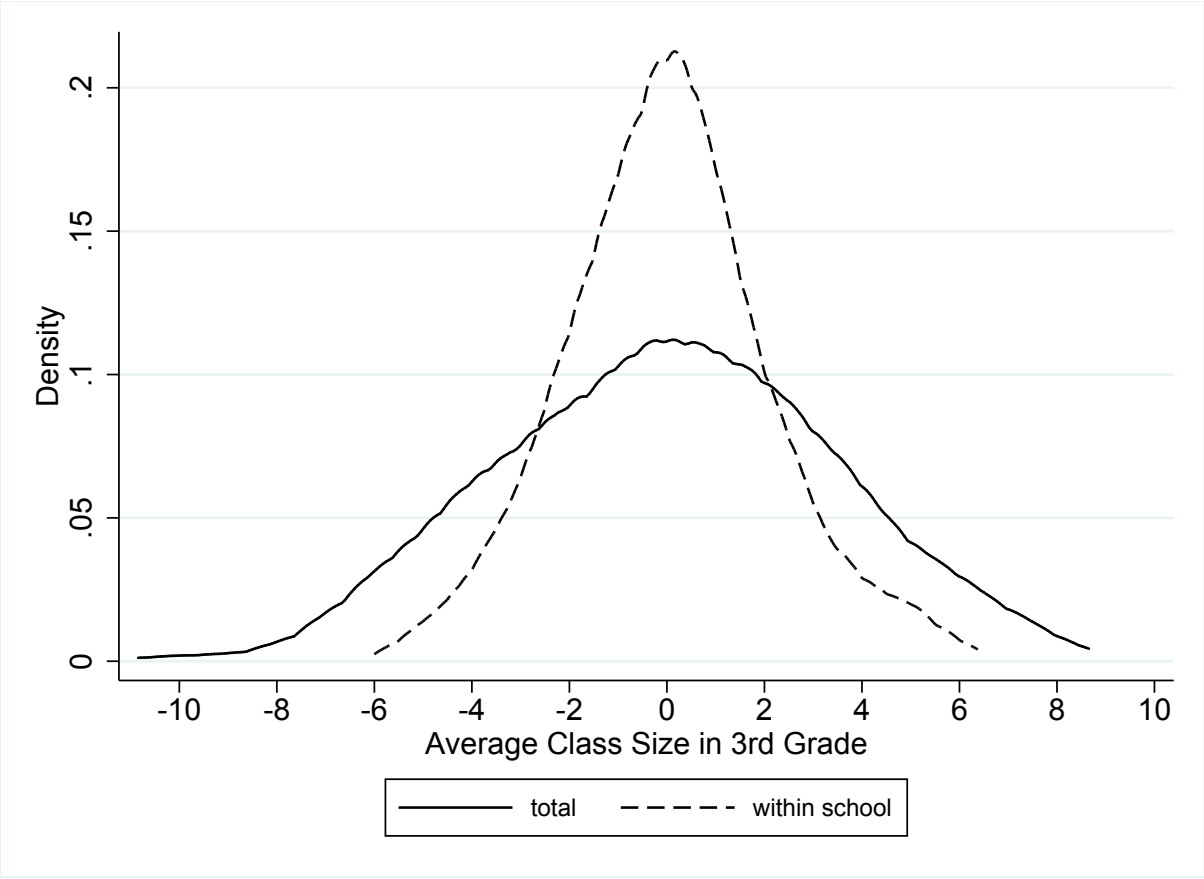
Table 11: The Effect of Class Size on Grade Repetition

	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: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Figures

Figure 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.

Appendix

A Figures and Tables

Table A.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 A.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.

Table A.3: Structure of NEPS Data

	2011	2012	2013	2013/2014	2014/2015	2015/2016
	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	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.

Table A.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: 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 number of classes. Columns 4-5 report instrumental variables estimates of average class size in grade 2 and 3 on the percentage of repeating-, late- and early enrolled students in grade 2 and 3. The instrument for class size the respective grade is imputed cohort size divided by number of classes. Regressions include school and year fixed effects. Standard errors clustered at the school-level are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.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: Each column contains results for a separate regression. Columns 1-4 report estimates of class size in grade 3 on language.

Table A.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

Table A.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: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.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 ≥ 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 ≥ 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: This table report 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.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: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.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 ≥ 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 ≥ 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: 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.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: This table reports IV results where each column in panels A and B contains the results for a separate regression with the same specification as that of column 6 in Table 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. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A.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 $\hat{\beta}$.

B Data

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.

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.[66] 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 A.3 shows when each test was conducted that goes 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.[67]

C Model Extensions

School System with Redshirting

Modifying our model to allow for reshirting corresponds to a simple relabeling of our model in section 3. LG now refers to the years in childcare before school entry and HG to the first grade in primary school. Children spend L years in childcare. 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 childcare, 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

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

[67] 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 must to be deduced in order to be able to select the right complement for the free field from the offered solutions.

effects of class size on academic skills in childcare and grade 1, respectively. The average test performance of students who were enrolled on time is then given in equation (10) and the average test performance of redshirted students is given in equation (11), where δ_s^t captures school and birth cohort-specific changes in skills associated with redshirting.

School System with Early Enrollment

To allow for early school enrollment in our model in section 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 3, the share of students from birth cohort t who enter grade 1 (HG) regularly in year $t+L$ is

$$\lambda_s^t = \frac{-\alpha_s^t + \theta + p_s^t}{2\theta} \tag{C.19}$$

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

$$N_{s\tau}^{obs} = \lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L+1}) N_s^{\tau-L+1} \tag{C.20}$$

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

$$\phi_s^\tau = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{N_{s\tau}^{obs}} = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{\lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L+1}) N_s^{\tau-L+1}} \tag{C.21}$$

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

$$E\left(test_{is}^{t}|regular\right) = E\left(test_{is}^{t}|test_{is}^{t} < p_{s}^{t}\right) = \frac{\alpha_{s}^{t} - \theta + p_{s}^{t}}{2} + \omega_{s,t+L} \tag{C.22}$$

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

$$\begin{aligned} E\left(test_{is}^{t}|early\right) &= E\left(test_{is}^{t}|test_{is}^{t} \geq p_{s}^{t}\right) + \delta_{s}^{t} + \omega_{s,t+L-1} \\ &= \frac{\alpha_{s}^{t} + \theta + p_{s}^{t}}{2} + \delta_{s}^{t} + \omega_{s,t+L-1} \end{aligned} \tag{C.23}$$

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

$$test_{s\tau} = \phi_{s}^{\tau-L} E\left(test_{is}^{\tau-L}|regular\right) + (1 - \phi_{s}^{\tau-L}) E\left(test_{is}^{\tau-L+1}|early\right) \tag{C.24}$$

D Proofs

To prove the results in section 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), $\Delta N_{s\tau}^{obs} = N_{s\tau}^{obs} - N_{s,\tau-1}^{obs}$, around $N_{s}^{t} = N$, $\alpha_{s}^{t} = \alpha$, and $p_{s}^{t} = p$ and we assume w.l.o.g. that $N = 1$. Making use of (6) and (7), this yields

$$\begin{aligned} \Delta N_{s\tau}^{obs} = &\left(\frac{\pi^{\alpha}}{2\theta} + \lambda\right) \Delta N_{s}^{\tau-L} + \left(1 - \lambda - \frac{\pi^{\alpha}}{2\theta}\right) \Delta N_{s}^{\tau-L-1} \\ &+ \frac{1}{2\theta}\left(\Delta\alpha_{s}^{\tau-L} - \Delta\alpha_{s}^{\tau-L-1} - \Delta p_{s}^{\tau-L} + \Delta p_{s}^{\tau-L-1}\right) \end{aligned} \tag{D.1}$$

where $\lambda = \frac{\alpha+\theta+p}{2\theta}$, $\Delta N_{s}^{t} = N_{s}^{t} - N_{s}^{t-1}$, $\Delta\alpha_{s}^{t} = \alpha_{s}^{t} - \alpha_{s}^{t-1}$ and $\Delta p_{s}^{t} = p_{s}^{t} - p_{s}^{t-1}$. Linearizing

the within-school change in the average test score in HG, $\Delta test_{s\tau} = test_{s\tau} - test_{s,\tau-1}$, using (2)-(12) yields

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

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

To prove the result in (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 cohort, $\alpha_s^t = \alpha$ and $p_s^t = p$. There are only shocks to cohort size as modeled in (2). In this case (D.1) and (D.2) simplify to

$$\Delta N_{s\tau}^{obs} = \lambda\Delta N_s^{\tau-L} + (1-\lambda)\,\Delta N_s^{\tau-L-1} \tag{D.3}$$

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

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

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

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

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

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 ($\lambda > 1/3$).

D.2 IV Results

To derive β_{IV} in (14), we need to calculate the covariances $Cov(\Delta test_{s,\tau}, \Delta N_{s\tau}^{obs})$ and $Cov(\Delta N_{s\tau}^{obs}, \Delta N_s^{\tau-1})$. Under our assumption of i.i.d. shocks to the cohort size N_s^t, η_s^t, it is straightforward to show

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

and

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

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

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

where

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

and

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

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.[68] This effect is likely to be very small relative to $3\lambda - 1$ and therefore can be neglected.[69] Using the same argument, it is easy to see from (D.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 (D.10), which include $\pi^{\alpha}/2\theta$, have only a negligible impact on the size of ρ_{IV}, which justifies the statement in footnote **??**.

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

To derive β_{IV}^{REA} in (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 (10) and (11). Therefore, the average test score in HG adjusted for the effect of grade retention at the individual level becomes

[68] To see this, simply take the derivative of $1 - \lambda_s^t$ with respect to N_s^t using (6).

[69] 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 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 2).

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

which differs from $test_{s\tau}$ in (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 (D.2) then yields

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

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

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

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

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

where ξ_{IV} is defined in (D.11).

D.3 OLS Results

To derive β_{OLS} in (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

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

and

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

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

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

where

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

and

$$\iota_{OLS} = \frac{\left(Var(\epsilon) - Var(\nu)\right)\frac{6\lambda-3}{4\theta} - \pi^{\omega}\frac{6}{4\theta^2}\left(Var(\epsilon) + Var(\nu)\right)}{Var(N_{s\tau}^{obs})} \tag{D.20}$$

and

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

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 (D.19) and (D.21) can be neglected. In that case, it is easy to show that ($\xi_{OLS} < 1$). The signs of (D.19) and (D.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.

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

Next, we derive β_{OLS}^{REA} in (17) following the same logic as in the previous two sections. The covariance of $\Delta test_{s\tau}^{REA}$ and $\Delta N_{s\tau}^{obs}$ can be shown to be

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

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

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

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

D.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 (13) under fairly general conditions. Theorem 1 summarizes this result:

Theorem 1 *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 1, let ϕ_h denote the autocovariance of η_s^t between year t and $t+h$. Using (D.3)-(D.4) and stationarity of η_s^t yields

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

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

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

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

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

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

To complete the proof, it remains to be shown that (D.27) is greater than (13) using conditions $(i) - (iv)$

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

E Simulation

We test our theoretical predictions by running simulations of a school systems 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 students 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:

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

- Class size is equal to

$$CS_{scg} = \frac{N_{scg}^{obs}}{C_{scg}}$$

- 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 on cohort and class size from these estimations can be found in Table A.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 3 based on actual data.