Online Appendix to "Birth Cohort Size Variation and the Estimation of Class Size Effects"

Maximilian Bach and Stephan Sievert

A Models of school systems with selective grade progression

A.1 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 (2014) below.[47] 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.[48] This helps to clarify what parameters are identified in different empirical designs.

In each year t a new cohort that consists of a continuum of students with mass N_s^t starts primary school in school s. To simplify the model, we assume that schools have only one class per grade, such that the number of students per grade and school corresponds to actual class size.[49] 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.

$$a_{is}^t > p_s^t \tag{A.6}$$

[47] Naturally, this section draws heavily on Ciccone and Garcia-Fontes (2014).

[48] Ciccone and Garcia-Fontes (2014) 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.

[49] 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 discuss the implications of class size thresholds in Section 5.

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

$$N_s^t = N_s + \eta_s^t \tag{A.7}$$

$$p_s^t = p_s + \nu_s^t \tag{A.8}$$

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 $Var(\nu_s^t) > 0$).[51] 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

$$\alpha_s^t = \alpha_s + \pi^{LG} N_s^t + \epsilon_s^t \tag{A.9}$$

where π^{LG} is the effect of class size in LG on academic skills and ϵ_s^t are i.i.d. shocks with mean zero and positive variance. In combination with the rule for grade retention in (A.6), this implies that the share of students (λ) in cohort t who are not retained and hence reach HG in year $t+L$ is[52]

$$\lambda_s^t = \frac{\alpha_s^t + \theta - p_s^t}{2\theta} \tag{A.10}$$

Class size in HG in school s in the school year starting in τ depends on the size of cohort

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

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

[52] 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:

$$-\theta \leq \alpha_s^t - p_s^t \leq \theta$$

$\tau - L$ and the share of nonretained students in that cohort as well as the size of cohort $\tau - L - 1$ and the share of retained students in that cohort

$$N_{s\tau}^{obs} = \lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L-1}) N_s^{\tau-L-1} \tag{A.11}$$

The share of nonretained students in HG in school s in the school year starting in τ is therefore

$$\phi_s^{\tau} = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{N_{s\tau}^{obs}} = \frac{\lambda_s^{\tau-L} N_s^{\tau-L}}{\lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L-1}) N_s^{\tau-L-1}} \tag{A.12}$$

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

$$\omega_{s\tau} = \tilde{\omega}_{s\tau} + \pi^{HG} N_{s\tau}^{obs} \tag{A.13}$$

where π^{HG} captures the effect of class size in HG and $\tilde{\omega}_{s\tau}$ are exogenous shocks. Thus, the sum $\pi^{LG} + \pi^{HG}$ captures the combined effect of class size in LG and HG on accumulated academic skills. This is our main parameter of interest, which we refer to as the "pure class size effect." At the end of HG, students take a standardized test. The average test performance of nonretained students reflects their academic skills accumulated in LG and HG, $a_{is}^t + \omega_{is,t+L}$. The average test performance of these students from cohort t who reach HG in year $\tau = t + L$ can be written as

$$E\left(test_{is}^t | nonretained\right) = E\left(test_{is}^t | a_{is}^t \geq p_s^t\right) = \frac{\alpha_s^t + \theta + p_s^t}{2} + \omega_{s,t+L} \tag{A.14}$$

where $E\left(a | a \geq p\right)$ denotes the average skills of nonretained 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

$$\begin{aligned} E\left(test_{is}^{t}|retained\right) &= E\left(test_{is}^{t}|a_{is}^{t} < p_{s}^{t}\right) \\ &= \frac{\alpha_{s}^{t} - \theta + p_{s}^{t}}{2} + \delta_{s}^{t} + \omega_{s,t+L+1} \end{aligned} \tag{A.15}$$

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 (A.12), (A.14) and (A.15)

$$test_{s\tau} = \phi_{s}^{\tau-L} E\left(test_{is}^{\tau-L}|nonretained\right) + (1 - \phi_{s}^{\tau-L}) E\left(test_{is}^{\tau-L-1}|retained\right) \tag{A.16}$$

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 the 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 A.3.

A.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.[53] The main reason is that within-school differences

[53] 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 (A.7).

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., $\pi^{LG} = \pi^{HG} = 0$). The instrumental variable approach exploiting variation in cohort sizes amounts to dividing the covariance of within-school changes of test scores in HG and within-school changes in cohort size by the covariance of within-school changes of cohort size in HG and initial cohort size. In Appendix D, we show that if there are no class size effects this ratio is equal to

$$\frac{3(\theta - \delta)(1 - \lambda)\lambda}{3\lambda - 1} \tag{A.17}$$

where $(\theta - \delta)$ is the average test score difference of nonretained students and students retained in the past, see (A.14) and (A.15), while λ is the average fraction of students who are not retained in LG. If $(\theta - \delta)$ is positive, i.e. nonretained 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 ($\lambda > 1/3$).

A.3 Model extensions

A.3.1 School system with redshirting

Modifying our model to allow for reshirting corresponds to a simple relabeling of our model in section A.1. 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 now refers to 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^{HG} capture the 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 (A.14) and the average test performance of redshirted students is given in equation (A.15), where δ_s^t captures school and birth cohort-specific changes in skills associated with redshirting.

A.3.2 School system with early enrollment

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

$$\lambda_s^t = \frac{-\alpha_s^t + \theta + p_s^t}{2\theta} \tag{A.18}$$

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

$$N_{s\tau}^{obs} = \lambda_s^{\tau-L} N_s^{\tau-L} + (1 - \lambda_s^{\tau-L+1}) N_s^{\tau-L+1} \tag{A.19}$$

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

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

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

$$E\left(test_{is}^t | regular\right) = E\left(test_{is}^t | test_{is}^t < p_s^t\right) = \frac{\alpha_s^t - \theta + p_s^t}{2} + \omega_{s,t+L} \tag{A.21}$$

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

$$\begin{aligned} E\left(test_{is}^t | early\right) &= E\left(test_{is}^t | test_{is}^t \geq p_s^t\right) \\ &= \frac{\alpha_s^t + \theta + p_s^t}{2} + \delta_s^t + \omega_{s,t+L-1} \end{aligned} \tag{A.22}$$

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

$$test_{s\tau} = \phi_s^{\tau-L} E\left(test_{is}^{\tau-L} | regular\right) + (1-\phi_s^{\tau-L}) E\left(test_{is}^{\tau-L+1} | early\right) \tag{A.23}$$

A.3.3 Implications

Analogous arguments to those in Section A.2 yield 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.

B Data

B.1 State-wide orientation exams Saarland

For 2003 and 2004, the development of test items for the centralized exams was carried out by the Bavarian State Institute of School Quality and Education Research, an organization with more than 50 years of experience in the field of educational consulting. In 2005 and 2006, this responsibility was transferred to Saarland's standing conferences on language and mathematics (Landesfachkonferenzen). Since the aim of the SOE was to safeguard quality assurance, test items were created such that they could assess students'

competences in relation to education standards set by the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder (Kultusministerkonferenz). The subject matter of the tests was the material from grades 2 and 3. In German, this related to the two domains of "Reading" and "Writing / Language and Use of Language." In reading, reference was made to the cognitive model of van Dijk and Kintsch (1983) that is also used in the international PIRLS studies. Questions were multiple choice and required extracting pieces of information from short texts. The most difficult questions further entailed meta-cognitive abilities, for example in the sense of relating texts to the author's likely intentions of writing them. In the domain of writing and use of language, spelling and grammar competences were specifically tested. Therefore, students had to complete words and reformulate sentences. The mathematics test was not further subdivided into different domains. However, all questions pertained to one or more of the following general mathematical competences: modelling, problem solving, argumentation, illustration, and communication. These competences had to be applied to specific mathematical content that students were supposed to be familiar with.

Standardized assessments may suffer from bias through manipulation of test scores by teachers (see, e.g., Angrist et al., 2017). 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, teachers did not correct the exams themselves. Answer sheet transcription and grading was performed by an external team of scorers who followed the provided grading rubrics. Potential test score manipulation by the teacher is thus unlikely.

B.1.1 Sample selection

We impose a set of restrictions on these data. First, we drop all schools for which we observe zero classes. These are schools that formed multi-grade classes because enrollment was too low to form separate classes for each grade. This restriction applies to 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. Note that we keep test scores for 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.

B.2 NEPS

The German National Education Panel Study (NEPS) was initially developed in 2009 to provide information on the determinants of education, the consequences of education, and to describe educational trajectories over the life course (Blossfeld, 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. 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 E.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

tests.[54]

C Additional results

C.1 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 student is retained in first, second, or third grade are 3.2%, 2.9%, and 2.8%, respectively. These are taken from from Table 1.
- 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

[54]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.

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 class size 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 C.1. 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 2 based on actual data.

C.2 Composition effect in Saxony

Here we replicate the results from Table 2 for another German federal state, Saxony. We have administrative, school-level enrollment and grade retention data for all public primary schools for the 2004-2015 school years. Columns 1-3 of Table C.2 show estimates for Saxony analogous to those reported in Table 2 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 regressed the fraction of students 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 2 by three.

C.3 Testing for random assignment of cohort size

Table C.3 tests whether student characteristics in grade 3 are balanced with respect to birth cohort size using the SOE student-level data. Each cell contains the result from a separate regression of the student characteristic listed in the leftmost column on the respective variable in the column head. 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 results from regressing the 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 prediction of a negative relationship between initial cohort size and the share of students held back or enrolled early on the grade-level, estimates for being older and younger than typical third graders

are significant and negative.[55] More generally, any significant effects in column 4 could be the result of compositional changes caused by initial cohort size. This can explain the significant negative coefficients for limited Germany proficiency and reporting none or few books at home as these are characteristics that correlate strongly with having been enrolled late or retained.

To test whether the initial birth cohort composition is balanced with respect to cohort size, we 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. Results are reported in column 5.[56] 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.[57]

C.4 Test Score differences between different groups of students

The theoretical results in Section 3 imply that instrumental variable estimates will be biased if nonretained students have skills that differ, on average, from retained, redshirted and early enrolled students. Here we test for average skill differences between these groups. As mentioned before, our data for Saxony only contain students' age in years. This precludes to distinguish between students who were enrolled one year too late and those who were retained in primary school, as they both are older than 9 years in our data. Further, we cannot distinguish between students who were enrolled one year early

[55] We suspect that these patterns were not discovered in previous within-school studies which performed similar balancing tests (e.g., Wößmann and West, 2006) because they only checked for linear relationships 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.

[56] 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. 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 thus 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.

[57] 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% 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%, only slightly above this expected value.

and those who were born between May and June but enrolled on time. Instead, we use data from the National Educational Panel Study (NEPS) starting cohort 2, which is a representative sample of primary school children from Germany. The NEPS contains standardized tests scores and information on whether a child has been retained and the timing of school enrollment. Thus, it allows identifying each group of students. Table C.4 reports results from regressions of language, math, and cognitive test scores 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 score 0.9 SD lower than regular students. Surprisingly, students who were enrolled early do not differ significantly from regular students. We can therefore expect the potential bias introduced by early enrollment to be of little concern.[58]

C.5 Testing for bias due to different class size thresholds

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 E.4, column 1 reports results where we regress the number of classes in grade 3 on an indicator for insufficient German proficiency measured in grade 3, total enrollment in grade 1, and school fixed effects. The positive coefficient for German proficiency indicates that grades with more students not proficient in German have significantly more classes holding enrollment constant. This, in turn, implies that class size for these students is about 0.169 students smaller than it is for students proficient in German from the same school with the same number of students in a grade; see column 2. Because of this feature of the data, we will control for German proficiency in some of the analyses below.

[58] Another potential concern are students who skip a grade. Table C.4 shows that these students score up to 0.96 SD higher 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 the NEPS data show that less than 0.6% of students skip a grade before grade 3 in Germany.

C.6 Alternative cohort-based approach for dealing with composition bias

An alternative approach to deal with bias arising from the mechanical correlation between cohort size and the grade-level composition of students in settings with grade repetition and redshirting is to reassign students to their original cohort.[59] To implement this approach, we reassign all students older than 9 on the day of testing to the previous cohort.[60] Note that the majority of students too old for their grade are either redshirted or repeat first grade (as opposed to second or third grade). Hence the class size of the cohort in which students are observed in grade 3 better reflects the class size they experienced up to grade 3. To decrease attenuation bias in IV estimates due to measurement error from assigning these students the "wrong" class size of their reassigned cohort, we only assign them the cohort identifier and cohort size but *not* the class size of the previous cohort.[61] The drawback of this approach, relative to directly controlling for age at test without reassignment, is a loss of observations as too old students from the first cohort cannot be reassigned.

Table C.5 presents IV estimates based on the reassignment approach for specifications with different sets of control variables in columns 5-7. For comparison, columns 1-4 reports estimates without reassignment from Table 5. Baseline estimates with reassignment but no further controls (column 5) are very similar to those with age controls but no reassignment (column 2). Adding further controls (excluding age) in columns 6-7 leads to similar coefficient movements as with age controls but no reassignment (columns

[59] We are grateful to one referee for suggesting this approach.

[60] We do not assign 8 year old students to the subsequent cohort because the majority of these students were not enrolled early or skipped a grade. Instead, approximately one sixth of our sample is 8 years old on the day of testing because tests were administered in May and the school enrollment cut-off is 30th of June. Regularly enrolled students born between May and June thus did not turn 9 yet on the day of testing. However, results are very similar when reassigning students younger than 9 to the subsequent cohort. This is to be expected given that early enrolled students, which make up the great majority of students too young for their grade, do not differ on average from regular students (see Table C.4).

[61] Attenuation bias occurs in the IV setting because the additional measurement error from assigning some students a class size, which they will not have experienced or only for one grade, results in a smaller reduced form that is not offset by a proportional decrease in the first stage. This is because, relative to no reassignment, the first stage will not change if we reassign both class and cohort size. The first stage in that case will not reflect the lack of correlation between reassigned cohort size and experienced class size for too old students and thus overstate the strength of the instrument.

3-4). This suggests that reassigning students too old for their grade to their initial cohort is similarly effective in reducing bias in IV class size estimates resulting from mechanical composition bias. However, given the loss of observations, estimates are less precisely estimated.

Table C.6 also reports corresponding OLS estimates where average class size is not instrumented. Columns 5-7 report estimates where older students are assigned the average class size of the previous cohort. Given the discussion above that the previous cohort's class size does not capture well experienced class size for students who have been redshirted or retained in first grade, reassignment leads to uniformly smaller class size estimates compared to no reassignment for OLS.

C.7 Effect heterogeneity

The specifications in Tables 5 and 6 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 this source of heterogeneity by interacting the class size variable with a set of indicator variables for being too old for grade 3, reporting few books at home, migration background, insufficient German proficiency, reading disorder (dyslexia), and learning disability in math (dyscalculia). We also test for heterogeneous effects by student gender. Tables C.7 and C.8 show estimates of these seven interactions with and without instrument, respectively. In line with the hypothesis that disadvantaged students are harmed most by larger classes, all interaction estimates in Table C.7, except for gender, are negative and most are statistically significant at the one percent level. IV estimates are similar but less precisely estimated. Additional evidence comes from the pattern of the interaction terms for dyslexia and dyscalculia. If students react more strongly to class size in subjects where they are at a disadvantage, we should expect larger effects for dyslexic students in language compared to math and vice versa for students with dyscalculia. This is exactly what we find in columns 6 and 7 in Panels A and B. Moreover, the interaction term for dyslexia is larger than the one for dyscalculia in language and vice versa in math, which we would also expect.

More importantly, the estimated class size effects for disadvantaged students are very large in magnitude: for example, the coefficient for insufficient German proficiency suggests that one more student in class decreases language and math test scores of students not proficient in German by 0.053 and 0.037 SD, respectively. Overall, these results reveal that our specifications in Tables 5 and 6 mask some marked effect heterogeneity for certain groups of students. Compared to 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 C.1: 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 $\hat{\beta}$.

Table C.2: 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	3,921	3,921	3,921	3,921	3,921

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

Source: Own calculations based on data from the Statistical Office of Saxony.

Table C.3: 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. Index refers to a

Table C.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	5,727	6,373	5,153

Notes: Each column reports estimates from a separate regression of the respective (standardized) test score on the variables listed in the rows. Robust standard errors are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Source: NEPS Data, Data Version SC2: 6.0.1.

Table C.5: IV Estimates of Class Size Effects on Test Scores: Controlling for Age versus Reassignment Approach

	IV				IV: Reassignment approach		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Language							
Class size	-0.0074	-0.0145*	-0.0189**	-0.0191**	-0.0139	-0.0189*	-0.0191**
	(0.0085)	(0.0085)	(0.0095)	(0.0092)	(0.0091)	(0.0099)	(0.0097)
N	37,847	37,847	37,847	37,847	36,164	36,164	36,164
Panel B: Math							
Class size	-0.0061	-0.0121	-0.0150	-0.0140	-0.0120	-0.0143	-0.0145
	(0.0108)	(0.0108)	(0.0111)	(0.0110)	(0.0112)	(0.0114)	(0.0114)
N	36,845	36,845	36,845	36,845	35,252	35,252	35,252
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls		Yes	Yes	Yes			
Insufficient German proficiency			Yes	Yes		Yes	Yes
Individual controls				Yes			Yes
N Cluster	156	156	156	156	156	156	156
N SchoolYearObs	828	828	828	828	828	828	828

Notes: Each cell contains results for a separate regressions of class size effects where class size in grade 3 is instrumented by predicted class size based on imputed cohort size. Columns 1-4 reports IV estimates from Table 5. Columns 5-7 report estimates where students older than 9 are assigned the cohort identifier and instrument of the previous cohort. Individual controls include dummies for gender, number of books at home, migration background, native language, and missing values for each variables. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Source: Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.

Table C.6: Estimates of Class Size Effects on Test Scores: Controlling for Age versus Reassignment Approach

	OLS				OLS: Reassignment approach		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Language							
Class size	-0.0159***	-0.0178***	-0.0202***	-0.0199***	-0.0124***	-0.0150***	-0.0151***
	(0.0045)	(0.0044)	(0.0052)	(0.0050)	(0.0042)	(0.0049)	(0.0047)
N	37,847	37,847	37,847	37,847	36,164	36,164	36,164
Panel B: Math							
Class size	-0.0112	-0.0127*	-0.0143**	-0.0140**	-0.0091	-0.0121*	-0.0123*
	(0.0068)	(0.0068)	(0.0072)	(0.0070)	(0.0069)	(0.0072)	(0.0070)
N	36,845	36,845	36,845	36,845	35,252	35,252	35,252
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age controls		Yes	Yes	Yes			
Insufficient German proficiency			Yes	Yes		Yes	Yes
Individual controls				Yes			Yes
N Cluster	156	156	156	156	156	156	156
N SchoolYearObs	828	828	828	828	828	828	828

Notes: Each cell contains results for a separate regressions of class size effects. Columns 5-7 report estimates where students older than 9 have been assigned the cohort identifier and class size of the previous cohort. Individual controls include dummies for gender, number of books at home, migration background, native language, and missing values for each variables. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Source: Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.

Table C.7: 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	36,845	36,845	36,845	36,845	36,845	36,845	36,845
	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	37,847	37,847	37,847	37,847	37,847	37,847	37,847
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 5, 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. Individual controls include dummies for age in years, gender, number of books at home, migration background, learnings disabilities, native language, and missing values for each variable. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. *Source:* Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.

Table C.8: 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	37,847	37,847	37,847	37,847	37,847	37,847	37,847
Cragg-Donald Wald F statistic	8,502	8,481	8,422	8,338	8,509	8,508	8,510
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	36,845	36,845	36,845	36,845	36,845	36,845	36,845
Cragg-Donald Wald F statistic	8,300	8,285	8,217	8,114	8,308	8,307	8,308
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 5, except that the class size variable is interacted with an indicator variable for the individual student characteristics. Individual controls include age in years, gender, number of books at home, migration background, and native language. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.
Source: Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.

D Proofs

To prove the results in Section 3 and Appendix A.2, 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 (A.10) and (A.11), this yields

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

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 (A.7)-(A.16) yields

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

D.1 Retention bias without "true class size effects"

To prove the result in (A.17), we assume that there are no class size effects, $\pi^{LG} = \pi^{HG} = 0$, 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 (A.7). 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 estimator 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 nonretained 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 (1), 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

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

and

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

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

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

where

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

and

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

ξ_{IV} will be approximately equal to 1/2. To see this note that $-\pi^{LG}/2\theta$ is the marginal effect of class size in LG on the share of grade repeaters in LG.[62] This effect is likely to be very small relative to $3\lambda - 1$ and therefore can be neglected.[63] Analogous arguments yield that the terms in (D.10), which include $\pi^{LG}/2\theta$, have only a negligible impact on the size of ρ_{IV}. It then follows that $\rho_{IV} \geq 0$ if class size has a negative effect on skills in LG, $\pi^{LG} < 0$ and the share of retained students is smaller than 1/3.

D.2.1 IV result controlling for the effect of grade retention at the individual level

To derive $\hat{\beta}_{IV}^{REA}$ in (2) 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 nonretained students in the same grade and school. This gap is $\theta - \delta$ (see, (A.14) and (A.15)). Therefore, the average test score in

[62] To see this, simply take the derivative of $1 - \lambda_s^t$ with respect to N_s^t using (A.10).

[63] 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 7). If we assume this effect is constant for grades 1 through 3, this estimate implies a value of $\pi^{LG}/2\theta$ 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 1).

HG adjusted for the effect of grade retention at the individual level becomes

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

which differs from $test_{s\tau}$ in (A.16) 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

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

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

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

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

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

where ξ_{IV} is defined in (D.11).

D.3 OLS results

To derive $\hat{\beta}_{OLS}$ in (3), 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

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

and

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

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

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

where

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

and

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

and

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

Using similar arguments about the relative magnitude of $\pi^{LG}/2\theta$ and λ as above, suggests that the terms involving $\pi^{LG}/2\theta$ in (D.19) and (D.21) can be neglected. In that case, it can be shown 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 (4) 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

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

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

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

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 spurious class size effect without any "true class size effects." For that case the spurious positive class size effect for the IV approach can be shown to be even larger than in the i.i.d. case in (A.17) 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) nonretained 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 (A.17) 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 Additional figures and tables

Table E.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 estimates that are significant at the 5 percent level. Level of data aggregation refers to the level at which the outcome variables are measured.

Table E.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 E.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 E.4: 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	38,415	38,415

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

Source: Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.

Table E.5: Estimates of Class Size Effects on Language Scores: Full Results

	IV				OLS			
	IV: Imputed cohort size				Avg. class size grade 3			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class size	-0.0074	-0.0145*	-0.0189**	-0.0191**	-0.0159***	-0.0178***	-0.0202***	-0.0199***
	(0.0085)	(0.0085)	(0.0095)	(0.0092)	(0.0045)	(0.0044)	(0.0052)	(0.0050)
Year 2004	-0.001	0.003	0.001	-0.458***	-0.003	0.002	0.001	-0.458***
	(0.025)	(0.024)	(0.026)	(0.054)	(0.025)	(0.024)	(0.026)	(0.054)
Year 2005	0.004	-0.024	-0.160***	-0.611***	0.016	-0.020	-0.158***	-0.610***
	(0.036)	(0.036)	(0.047)	(0.063)	(0.035)	(0.035)	(0.045)	(0.061)
Year 2006	-0.005	-0.028	-0.158***	-0.576***	0.004	-0.025	-0.157***	-0.575***
	(0.033)	(0.034)	(0.041)	(0.058)	(0.033)	(0.033)	(0.040)	(0.057)
Younger than 9 at test	—	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 years old at test	—	-0.755***	-0.495***	-0.451***	—	-0.755***	-0.495***	-0.451***
		(0.022)	(0.020)	(0.019)		(0.022)	(0.020)	(0.019)
11 years old at test	—	-1.030***	-0.668***	-0.576***	—	-1.030***	-0.668***	-0.576***
		(0.052)	(0.048)	(0.046)		(0.052)	(0.048)	(0.046)
Age missing	—	-0.306***	-0.279**	-0.088	—	-0.305***	-0.279**	-0.087
		(0.102)	(0.110)	(0.208)		(0.102)	(0.110)	(0.208)
Insufficient German proficiency	—	—	-0.910***	-0.834***	—	—	-0.910***	-0.835***
			(0.016)	(0.015)			(0.016)	(0.015)
Insufficient German proficiency missing	—	—	-0.389***	-0.374***	—	—	-0.389***	-0.374***
			(0.047)	(0.046)			(0.047)	(0.046)
Male	—	—	—	-0.136***	—	—	—	-0.136***
				(0.009)				(0.009)
Male missing	—	—	—	-0.194	—	—	—	-0.195
				(0.179)				(0.179)
Book: Enough to fill one shelf	—	—	—	0.206***	—	—	—	0.206***
				(0.028)				(0.028)
Books: Enough to fill one bookcase	—	—	—	0.340***	—	—	—	0.340***
				(0.026)				(0.026)
Books: Enough to fill two bookcases	—	—	—	0.405***	—	—	—	0.405***
				(0.026)				(0.026)
Books: more than 200	—	—	—	0.475***	—	—	—	0.475***
				(0.028)				(0.028)
99.booksIM	—	—	—	-0.112**	—	—	—	-0.112**
				(0.054)				(0.054)
Migration background	—	—	—	-0.060	—	—	—	-0.060
				(0.037)				(0.037)
Migration background missing	—	—	—	-0.198**	—	—	—	-0.197**
				(0.077)				(0.077)
Non-native German speaker	—	—	—	-0.075**	—	—	—	-0.075**
				(0.032)				(0.032)
Non-native German speaker missing	—	—	—	0.115	—	—	—	0.114
				(0.094)				(0.094)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	37,847	37,847	37,847	37,847	37,847	37,847	37,847	37,847

Notes: Each column contains results for a separate regression. Columns 1-4 report estimates of class size in grade 3 on language where class size is instrumented by predicted class size based on imputed cohort size. Columns 5-8 report estimates of class size in grade 3 on language. Individual controls include dummies for gender, number of books at home, migration background, native language, and missing values for each variable. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. *Source:* Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.

Table E.6: Estimates of Class Size Effects on Math Test Scores: Full Results

	IV				OLS			
	IV: Imputed cohort size				Avg. class size grade 3			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class size	-0.0061	-0.0121	-0.0150	-0.0140	-0.0112	-0.0127*	-0.0143**	-0.0140**
	(0.0108)	(0.0108)	(0.0111)	(0.0110)	(0.0068)	(0.0068)	(0.0072)	(0.0070)
Year 2004	-0.002	0.000	-0.000	-0.320***	-0.003	0.000	-0.000	-0.320***
	(0.032)	(0.032)	(0.034)	(0.047)	(0.032)	(0.032)	(0.034)	(0.047)
Year 2005	-0.034	-0.056	-0.154***	-0.469***	-0.027	-0.056	-0.155***	-0.469***
	(0.047)	(0.047)	(0.051)	(0.062)	(0.045)	(0.045)	(0.049)	(0.060)
Year 2006	-0.042	-0.059	-0.154***	-0.441***	-0.037	-0.059	-0.154***	-0.441***
	(0.047)	(0.047)	(0.050)	(0.062)	(0.046)	(0.046)	(0.048)	(0.061)
Younger than 9 at test	—	0.079***	0.051***	0.051***	—	0.079***	0.051***	0.051***
		(0.016)	(0.015)	(0.015)		(0.016)	(0.015)	(0.015)
10 years old at test	—	-0.612***	-0.420***	-0.403***	—	-0.612***	-0.420***	-0.403***
		(0.022)	(0.021)	(0.020)		(0.023)	(0.021)	(0.020)
11 years old at test	—	-0.763***	-0.499***	-0.463***	—	-0.763***	-0.499***	-0.463***
		(0.047)	(0.045)	(0.044)		(0.047)	(0.045)	(0.044)
Age missing	—	-0.249**	-0.258**	0.046	—	-0.249**	-0.257**	0.046
		(0.124)	(0.128)	(0.190)		(0.124)	(0.128)	(0.190)
Insufficient German proficiency	—	—	-0.669***	-0.655***	—	—	-0.669***	-0.655***
			(0.017)	(0.017)			(0.017)	(0.017)
Insufficient German proficiency missing	—	—	-0.254***	-0.238***	—	—	-0.255***	-0.238***
			(0.053)	(0.054)			(0.053)	(0.054)
Male	—	—	—	0.204***	—	—	—	0.204***
				(0.009)				(0.009)
Male missing	—	—	—	-0.140	—	—	—	-0.140
				(0.144)				(0.144)
Book: Enough to fill one shelf	—	—	—	0.182***	—	—	—	0.182***
				(0.030)				(0.030)
Books: Enough to fill one bookcase	—	—	—	0.322***	—	—	—	0.322***
				(0.031)				(0.031)
Books: Enough to fill two bookcases	—	—	—	0.374***	—	—	—	0.374***
				(0.033)				(0.033)
Books: more than 200	—	—	—	0.441***	—	—	—	0.441***
				(0.034)				(0.034)
99.booksIM	—	—	—	0.010	—	—	—	0.010
				(0.050)				(0.050)
Migration background	—	—	—	0.024	—	—	—	0.024
				(0.044)				(0.044)
Migration background missing	—	—	—	-0.118	—	—	—	-0.118
				(0.073)				(0.072)
Non-native German speaker	—	—	—	0.005	—	—	—	0.005
				(0.037)				(0.038)
Non-native German speaker missing	—	—	—	0.034	—	—	—	0.034
				(0.106)				(0.106)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	36,845	36,845	36,845	36,845	36,845	36,845	36,845	36,845

Notes: Each column contains results for a separate regression. Columns 1-4 report estimates of class size in grade 3 on math where class size is instrumented by predicted class size based on imputed cohort size. Columns 5-8 report estimates of class size in grade 3 on math. Individual controls include dummies for gender, number of books at home, migration background, native language, and missing values for each variable. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance

Table E.7: The Effect of Class Size in Different Grades on Test Scores

	IV				OLS			
	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	-0.0140**	-0.0171**	-0.0191**	-0.0160**	-0.0109**	-0.0105**	-0.0199***	-0.0153***
	(0.0068)	(0.0080)	(0.0092)	(0.0077)	(0.0055)	(0.0050)	(0.0050)	(0.0054)
Math	-0.0102	-0.0123	-0.0140	-0.0117	-0.0095	-0.0061	-0.0140**	-0.0109
	(0.0080)	(0.0095)	(0.0110)	(0.0092)	(0.0068)	(0.0067)	(0.0070)	(0.0074)
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 where class size is instrumented by predicted class size based on imputed cohort size. Columns 5-8 report estimates of class size in different grades on language and math. Individual controls include gender, number of books at home, migration background and native language. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$. *Source:* Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.

Table E.8: Robustness Checks: Different Specifications

	IV			OLS		
	(1)	(2)	(3)	(4)	(5)	(6)
Language	-0.019**	-0.031	-0.016	-0.020***	-0.027***	-0.020***
	(0.009)	(0.020)	(0.015)	(0.005)	(0.010)	(0.007)
N	37,847	15,386	37,847	37,847	15,386	37,847
Cragg-Donald Wald F statistic	17,017	4,484	11,648			
Kleibergen-Paap rk Wald F statistic	176.48	38.42	86.29			
Math	-0.014	-0.041	-0.021	-0.014**	-0.019	-0.021**
	(0.011)	(0.026)	(0.018)	(0.007)	(0.012)	(0.009)
N	36,845	14,944	36,845	36,845	14,944	36,845
Cragg-Donald Wald F statistic	16,614	4,366	11,304			
Kleibergen-Paap rk Wald F statistic	175.77	38.05	84.89			
School FE	Yes	Yes	Yes	Yes	Yes	Yes
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-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 where class size is instrumented by predicted class size based on imputed cohort size. Columns 5-8 report estimates of class size in grade 3 on language and math. Individual controls include gender, number of books at home, migration background, and native language. Standard errors clustered at the level of the combined schools in 2005 are given in parentheses. Significance level: * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.
Source: Own calculations based on SOE waves 2003-2006 and data from the Statistical Office of Saarland.