Astrid Marie Jorde Sandsør

Educational Policy and Student Outcomes

February, 2016

Dissertation for the Ph.D. degree
Department of Economics
University of Oslo

To my daugher, Marie

Acknowledgements

This thesis has been written while being employed at Department of Economics at the University of Oslo. I have been associated with the Centre for the Study of Equality, Social Organization and Performance (ESOP) at the Department of Economics, University of Oslo. ESOP is supported by the Research Council of Norway.

I would like to thank my supervisors Kalle Moene and Torberg Falch for excellent supervision and Nina Drange, Tarjei Havnes, Torberg Falch and Bjarne Strøm for excellent research collaboration. Finally, I would like to thank my family for their support and my fellow Ph.D. students for making every day better, especially Tord, André, Esther Ann, Åshild, Kari, Anna, Kristin and Nina, not to mention Lasse who also became my husband.

Contents

Introduction

Economics is concerned with "the allocation of scarce means to satisfy competing ends" (Becker, 1978, p.3). A country has limited resources to distribute between services such as health care and schooling, and society must strive to spend these resources in the best way possible. With each sector competing for limited resources and various policies to chose from, we need measures to determine how well our resources are spent. Should we spend additional resources to increase life expectancy or to increase learning? If we want to increase learning, should we do so by reducing class size or starting school at an earlier age? If learning is our goal, how do we measure learning?

Measuring increases in human capital, a person's knowledge or skills, is one way of measuring efficient resource use in education (Schultz, 1961; Mincer, 1970; Becker, 1964). The Mincer equation (Mincer, 1970, 1974) models income as a function of human capital, defined in terms of years of education and potential labor market experience. Recent availability of data has made it possible to expand on the measure of human capital to include measures of quality rather than just quantity. Hanushek and Woessmann (2011) consider cognitive skills, identified by test scores, as good measures of relevant skills for human capital, while other studies have used school grades as measures of cognitive ability (Falch, Nyhus, and Strøm, 2014a,b; Leuven, Oosterbeek, and Rønning, 2008). Cognitive ability has been shown to be an important predictor for future outcomes for the individual, including education and labor market outcomes (Murnane, Willett, and Levy, 1995; Herrnstein and Murray, 2010; Heckman, 1995), and aggregate measures of cognitive abilities are important for economic growth and development (Hanushek and Woessmann, 2008; Hanushek and Kimko, 2000).

An emerging literature focuses on the importance of non-cognitive skills for human capital. Non-cognitive skills are skills such as perseverance, conscientiousness, self-control, trust, attentiveness, self-esteem and self-efficacy, resilience to adversity, openness to experience, empathy, humility, tolerance of diverse opinions and the ability to engage productively in society (Kautz, Heckman, Diris, ter Weel, and Borghans, 2014, p. 9), and have been shown to be meaningful predictors of educational, labor market and behavioral outcomes (Kautz, Heckman, Diris, ter Weel, and Borghans, 2014; Heckman, Stixrud, and Urzua, 2006; Borghans, Duckworth, Heckman, and Ter Weel, 2008; Carneiro, Crawford, and Goodman, 2007; Falch, Nyhus, and Strøm, 2014b; Lindqvist and Vestman, 2011). More importantly, studies have shown that non-cognitive skills are malleable and are dynamically related to

cognitive skills, such that boosting non-cognitive skills early in life can increase the benefits of education later in life (Cunha and Heckman, 2007).

Human capital, measured by cognitive or non-cognitive skills, is important for the individual and for society, so what can policy makers do to increase cognitive and non-cognitive skills? Individual outcomes, such as personality tests, academic test scores or years of education, are often modeled as functions of student characteristics, family background, resources, institutional features and individual ability (Hanushek and Woessmann, 2011). We are interested in the factors that we can be changed though policy. We can change resources by increasing investments during childhood or schooling, or we can change the institutional features of early childhood, the educational system or the labor market.

One reason to focus on investments during early childhood is that returns are likely to be high. There is more time to reap the rewards (Becker, 1964) and investments in human capital have dynamic complementarities (Cunha and Heckman, 2007). Currie (2001) argues that governments concerned with equity should attempt to equalize initial endowments through early childhood education rather than compensating for differences later on in life. In addition, studies in neuroscience and developmental psychology indicate that learning is easier in early childhood than later in life (Shonkoff, Phillips, and Council, 2000). Empirical evidence suggests that preschool programs, especially those aimed at disadvantaged children, can have both short and long term benefits (Almond and Currie, 2011; Knudsen, Heckman, Cameron, and Shonkoff, 2006; Ruhm and Waldfogel, 2012; Baker, 2011), as can starting school at an earlier age (Black, Devereux, and Salvanes, 2011; Leuven, Lindahl, Oosterbeek, and Webbink, 2010). However, other studies find no or even negative effects of preschool programs and early enrollment into school (Cascio, 2009; Gupta and Simonsen, 2010; Drange, Havnes, and Sandsør, 2012; Baker, Gruber, and Milligan, 2008).

We could also chose to increase resources in schooling. Ever since the Coleman Report (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York, 1966) presented evidence often interpreted as school resources being unimportant for student performance, researchers have been trying to investigate the role and impact of resources in school. Increasing the teacher-student ratio (either more teachers or smaller classes), hiring better teachers or improving the school facilities are all ways of increasing school resources. Since then, both the availability of data and methods to uncover causal effects have increased, but results remain inconclusive (Hanushek, 1986, 2003, 2006; Webbink, 2005). Even for the narrow and popular policy tools of reducing class size, studies differ substantially in their conclusions.

While studies from the famous randomized experiment in Tennessee (STAR) find both short and long term positive effects of reduced class size (Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011; Dynarski, Hyman, and Schanzenbach, 2013; Krueger and Whitmore, 2001), results using quasi-experimental methods are mixed (Angrist and Lavy, 1999; Hoxby, 2000; Wößmann and West, 2006; Fredriksson, Öckert, and Oosterbeek, 2013; Leuven, Oosterbeek, and Rønning, 2008; Falch, Strøm, and Sandsør, 2015), suggesting that the effects of reduced class size are context dependent.

Improving cognitive and non-cognitive outcomes can also be achieved through policies that change institutional features. Changing how schools are organized or changing the curriculum are both examples of changing institutional features without necessarily increasing resources. Another way of changing institutional features is to change the size of the school district, however it is not clear whether larger school districts are associated with improved outcomes or the opposite. On the one hand, there could be economies of scale associated with district size such that larger districts provide resources to students more efficiently than smaller districts. For example, the probability of hiring professional and able school administrators may be higher in large than in small school districts. On the other hand, increasing district size might decrease local autonomy, and if the student population becomes more heterogeneous as a result, the larger district might be less able to meet the needs of the students. Results from studies using quasi-experimental methods to investigate the effect of school district size on student outcomes are mixed (Gordon and Knight, 2008; Berry and West, 2010; Beuchert, Humlum, Nielsen, and Smith, 2015; Reingewertz, 2012).

Another institutional feature that can be changed is how acceptance into further education is determined. Admission could be determined by for instance proximity to the school or ability or a combination of criteria. When applying to higher education in the United States, the major determinants for admission are grades in college preparatory courses, test scores from the ACT or SAT, and overall grades. Class rank, an application essay or writing samples and letters of recommendation may also be admission criteria (Clinedinst and Hawkins, 2011). In Norway, however, students apply to higher education almost entirely based on their grade point average from upper secondary education. Admission systems have two functions. First, they affect student incentives. Haraldsvik (2012), for instance, finds that the introduction of free school choice in publicly provided upper secondary education in Norway increased student performance in lower secondary education. Second, they try to achieve the best match between institutions and students. If institutions are

interested in students with high ability and effort, but are not able to measure this optimally, they may not be accepting the best students.

There are few agreed upon truths in the field of education. We know education is important, and that teachers matter, but we know less about how to improve education or teacher quality. Many papers are correlational studies that do not lend themselves to conclusions about causality, and even when experimental or quasi-experimental methods are used, results differ depending on country and context. The chapters in this thesis are no exception.

The first two chapters of this thesis contributes to the literature on the effect of increased resources on student performance in the long run. Although there are good reasons to believe that early childhood investments are important for both short-run and long-run outcomes, there are studies finding both positive, no or even negative effects. In chapter one, we study a Norwegian reform mandating kindergarten at age 5-6, finding no long-run effects on educational outcomes.

Project STAR found important and significant effects of reduced class size, but the evidence from quasi-experimental studies is mixed. In some countries and contexts, smaller classes improve outcomes, but this does not seem to be the case for Norway. In chapter two, we study the effect of class size in Norway on long-run outcomes, educational attainment and income in adulthood using a quasi-experimental design and find no effect of changing class size.

Chapter three of the thesis contribute to the literature on how institutional features affect student outcomes. Theoretically, it is not obvious whether larger or smaller school districts should lead to better student outcomes, and empirically the results are mixed. We study the effect of changing school district size in Norway on income in adulthood and find a positive effect.

Chapter four contributes to the literature on measuring human capital by studying the impact of a new measure of skill; individual grade variance. Grade variance is found to be negatively associated with educational attainment across the grading distribution. If institutions are using grade point average as their main determinant of admission, then students with low grade variance who are just below the grade point average cutoff are likely to outperform student with high grade variance just above the cutoff. This finding suggests that institutions should take other measures of ability into account in the admission decisions.

If our goal is to improve education, factors tied to successful policy interventions need to be identified. This can only be done by conducting rigorous studies in different contexts. Only then can we know that our limited resources in education are well spent.

Chapter 1: Kindergarten for all: Long-run effects of a universal intervention

Nina Drange, Tarjei Havnes and Astrid Marie Jorde Sandsør

Universally available child care of high quality can benefit child development, also in the long run (Almond and Currie, 2011). Returns are often found to be particularly high for children from disadvantaged families. At the same time, children from disadvantaged families are underrepresented in existing programs. This sorting into the programs coupled with particularly large estimated benefits among disadvantaged children, suggests a potentially strong social gradient in expanding or mandating early childhood interventions (Barnett and Belfield, 2006). Indeed, in an effort to counter differences at school entry depending on social background, many countries are currently moving towards subsidized child care available for the general population.

Policies and proposals promoting universal interventions in early childhood pose a challenge to the existing literature, which has reserved most of its attention for programs targeted at disadvantaged children. Existing studies on universally available programs typically reveal the impact on children from families with a strong preference for out-of-home care. For instance, Baker, Gruber, and Milligan (2008) and Havnes and Mogstad (2011) study the introduction of a universally available program but with actual enrollment being far from universal, while Gupta and Simonsen (2010) explicitly exploit rationing of child care for identification. Since both theory and evidence point towards important heterogeneity in the effects of early childhood interventions, it remains an open question how well the current evidence can inform about the impact of truly universal interventions. In particular, it is unclear how effective programs with universal participation may be at addressing the needs of disadvantaged children.

In the current paper, we provide evidence on the long-run effect on schooling of a Norwegian reform that mandated kindergarten at age 5–6. We first consider the impact on children's school performance at the end of compulsory schooling at age 15–16. We also consider the impact on high school dropout (age 18) and on enrollment in the academic track in upper secondary school (age 16). These are interesting in their own right, and help confront the concern of fading out of cognitive effects from early intervention programs, even when long-term effects on substantive outcomes may persist. Our identifying variation comes from a 1997-reform in Norway that lowered school starting age from seven to six. The new program for six year olds was designed as a low intensity kindergarten program, aimed at prepar-

ing children for school by learning through play, similar to early U.S. kindergarten programs (Cascio, 2009). The goal of the new program was to counter differences in learning outcomes between children from different socioeconomic backgrounds. While disadvantaged children were thought to benefit most from kindergarten programs, they were strongly underrepresented in the existing voluntary programs prior to the reform.

As the implementation of the reform was nationwide, the most direct assessment compares cohorts just young enough to be affected with cohorts just old enough not to be affected. An immediate objection to this strategy is that we may be confounding effects of the policy with unrelated cohort effects. To get around this issue, we take advantage of voluntary enrollment in child care prior to the reform. Since the new program for six year olds bears strong resemblance to kindergarten programs that were widely available prior to the reform, it should have little impact on children that would voluntarily enroll in such programs. This motivates a difference-in-differences approach where we compare outcomes before and after the implementation of the mandatory kindergarten reform, of children who enroll in voluntary kindergarten at age six (i.e. the control group) and children who do not enroll in voluntary kindergarten at age six (i.e. the treatment group). As voluntary enrollment at age six is unobserved by definition after the reform, we use enrollment in child care at age five to determine treatment. This should be a good proxy since children who are enrolled at age five are almost universally enrolled at age six.

Results reveal that the program had little impact on affected children. In our baseline estimation, the precisely estimated effect on the child's school performance is negative but below 2 % of a standard deviation. Meanwhile, we find a modest increase in high school dropout rates, and no impact on academic tracking in upper secondary school. These results are robust to including or excluding a large set of observable characteristics, as well as a battery of specification checks confronting the key identifying assumption of common trends in treatment and comparison groups. Importantly, we find no evidence of a separate effect of the reform on our comparison group that may attenuate effects on the treatment group, nor of a delayed effect of the reform on later cohorts. We also find no evidence of important heterogeneity when we look across subsamples reflecting the child's background and home environment, across different segments of the grading distribution, or across school subjects where we may expect children to benefit from different types of skills.

Our paper contributes to the rapidly increasing literature on how early childhood interventions in general, and kindergarten programs in particular, can promote the formation of skills in children. This literature is divided into two distinct branches,

one focused on targeted programs, the other focused on universal programs available to the general population. While studies of targeted programs often find positive effects, the literature on universal programs is smaller and findings are mixed. Perhaps as a consequence, the discussion on child care policies is based largely on the targeted literature and descriptive evidence, even when the policies considered are universal.

The current paper contributes to the literature on universal child care programs in two distinct ways. First, since kindergarten is not rationed prior to mandating, the estimated effect should derive from the particular group of children that do not voluntarily enroll. Our study therefore provides a rare opportunity to learn about the group of never-takers, to use the terminology of Imbens and Angrist (1994). This is of particular interest since these families may have quite different characteristics compared to families that select into child care voluntarily, many of which may be unobserved. If so, then existing estimates may tell us little about the potential effect of child care among these children. Second, while the program we study is universal, the reform may be viewed as targeted since affected children come disproportionately from disadvantaged families. Our results may therefore shed light on how a universal low intensity program can improve outcomes among the disadvantaged. That is, can the positive effects for the disadvantaged, often seen from targeted interventions, be reproduced in a universal program? Our evidence suggests that this is not the case. This is true, even though the estimates likely reflect shifts mostly from parental care, rather than informal care.

We believe that our evidence may call for caution in the current push towards using universal child care as a tool to promote the development of children from disadvantaged families. While we agree that early childhood investments can be an important tool in facilitating equal opportunities, our evidence emphasizes that this is hardly automatic, and suggests that the structuring of the program and its content may be key to generating the intended benefits.

Chapter 2: Do smaller classes always improve students' long-run outcomes?

Torberg Falch, Astrid Marie Jorde Sandsør and Bjarne Strøm

The impact of school resources on student performance has been disputed since the publication of the Coleman Report (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York, 1966). Although availability of data and empirical strategies to uncover causal effects have increased substantially in recent years, the evi-

dence on the effect of resources on education outcomes is still inconclusive. The literature is not conclusive even for more narrow and popular policy tools as class size. Although the results from the well known randomized Student/Teacher Achievement Ratio experiment (Project STAR) in Tennessee (Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011; Krueger and Whitmore, 2001) suggest that smaller classes are beneficial in terms of test scores, studies using quasi-experimental approaches to identify causal effects differ substantially in their conclusions. One interpretation is that extra resources and reduced class size are effective tools in some contexts, while ineffective in other contexts.

Academic test scores only measure cognitive skills, while class size may also affect non-cognitive skills. In addition, evidence based on test scores may be biased in settings where teachers systematically manipulate test scores as recently demonstrated in Angrist, Battistin, and Vuri (2015). Both arguments suggest that analyses of long-run outcomes in terms of educational attainment and income in adulthood as used in our empirical study would provide the most credible evidence of the effect of school resources. Such studies will embed all short-run effects in addition to effects on non-cognitive skills that are difficult to measure directly.

Three recently published papers analyze long-run effects of class size. Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) and Dynarski, Hyman, and Schanzenbach (2013) study long-run outcomes for participants in the STAR experiment, while Fredriksson, Öckert, and Oosterbeek (2013) exploit a class size rule in Sweden to estimate both short-run and long-run outcomes. These papers all find positive long-run effects of smaller classes, suggesting that the mixed effects in the literature on short-run effects are related to imperfect measurement of student skills. However, the findings for the long run are also consistent with the findings in the short run using test scores within the same contexts. Of particular interest is Fredriksson, Öckert, and Oosterbeek (2013) who find a positive short-run effect on non-cognitive ability; an outcome rarely available for researchers. These results motivate studies on long-run outcomes from contexts where the evidence indicates no class size effect on short-run outcomes.

In this paper we estimate long-run effects of class size for Norway where previous research has not been able to provide evidence of short-run gains from smaller classes in terms of student achievement. We investigate whether the class size effect in lower secondary education depends on characteristics of the environment in which the schools and students operate. Leuven and Løkken (2015) explore similar data, estimating the effect of class size both in primary and lower secondary education. Their analysis utilizes that some schools include grades 1 to 10, assuming that

the students stayed in the same school during all school years. We find qualitatively similar effects of class size as they do and extend the analysis to investigate potential heterogeneous effects across school districts.

The findings for short-run outcomes differ substantially between the Scandinavian countries Sweden, Denmark and Norway with apparently similar educational and labor market institutions. All countries have small income differences, generous welfare state arrangements, and comprehensive public school systems seeking to equalize opportunities across families and students. Nevertheless, closer inspection reveals that important institutional differences prevail with regard to for instance school district size and teacher shortages.

We first exploit the strict class size rule in Norway and match individual and school register information from 1982 through 2011 to estimate causal effects on educational attainment and income. While experimental studies are often viewed as the "gold standard" in empirical research, exploiting the class size rule in a quasi-experimental approach makes it possible to circumvent the potential Hawthorn effect that might plague experimental studies (Ehrenberg, Brewer, Gamoran, and Willms, 2001). In contrast to Fredriksson, Öckert, and Oosterbeek (2013), we are able to use register data for the whole population of schools for cohorts born 1966-1984 representing almost 1 million students and 1150 schools with separate catchment areas.

Secondly, information on the whole population of schools and students offers a unique possibility to use the quasi-experimental strategy to study whether the class size effect depends on characteristics of the environment in which the schools and students operate. We focus on dimensions that mirror differences in external conditions indicated by previous studies to be important for school efficiency and student performance, such as teacher quality, extent of upper secondary school choice, school district size, local fiscal constraints and labor market conditions.

We find insignificant effects of class size in grades 8-10 on educational attainment and income. While this is in contrast to the previous papers on long-run effects, it is in accordance with the findings in the short run for Norway and the long-run effect in Leuven and Løkken (2015). Moreover, we find no evidence that class size effects vary with school district characteristics.

Chapter 3: Municipality mergers

Astrid Marie Jorde Sandsør and Bjarne Strøm

The size and number of local governments is an important policy question. Municipal amalgamation reforms and consolidation of school districts have emerged in many countries and the issue is currently on the political agenda in countries like Norway and Finland. While fiscal decentralization is generally believed to be beneficial for society as suggested by the decentralization theorem formulated by Oates (1972), common arguments for amalgamation reforms are that larger units realize economics of scale. According to this argument, increased school district size implies reduced expenditure per pupil. However, the size effect on output quality is not obvious. Expenditure reduction may come at the cost of reduced quality of services provided by the local units. On the one hand, larger local units may decrease local autonomy at the provider level (school, day care institution or homes for elderly). If the population becomes more heterogeneous as a result, the larger local governments might be less able to meet the needs of the heterogeneous users of public services. On the other hand, it is possible that larger local governments will have more professional administration and management of resources and so increase output quality for a given amount of resources available. For example, the probability of hiring professional and able school administrators may be higher in large than in small school districts. Ultimately, the relationship between local government size and output quality can only be resolved by empirical studies.

This paper contributes to the literature by investigating the effect of municipal size on educational output in terms of student educational attainment and earnings in adulthood using rich data from administrative registers in Norway. In order to provide credible evidence, we explore the spatial and temporal variation in municipal size from enforced municipality mergers taking place in Norway in the 1980's and 1990's in a difference-in-differences approach. Using outcomes in terms of educational attainment and earnings has several advantages when studying the relationship between municipality size and output quality. First, educational services in terms of compulsory schooling is provided by all municipalities, small and large. The users are well defined (children age 7-16) and to the extent that private schooling is not an option, services are solely provided by the local public sector. Second, educational attainment and earnings in adulthood may be more relevant measures of education output than test scores often used in estimates of education production functions as these broader measures are more likely to reflect the multi-dimensional property of educational production. Third, we can control for individual socioeco-

nomic characteristics in the analysis. Lastly, we are able to use a school fixed effects strategy. To the extent that municipality mergers did not lead to school consolidation, we can compare students before and after the merger attending the same schools.

The mergers we study were enforced by the central authorities based on recommendations from two official Norwegian reports (Norwegian Ministry of Local Government and Labor, 1986, 1989). The mergers were former city municipalities merging with surrounding municipalities, having two main benefits. First, it creates a natural comparison group of city and surrounding municipalities. Second, there is reason to believe that merging could have different consequences for the city and surrounding municipalities. The mergers were often met by large local resistance in the municipalities surrounding the city and several referenda gave very little support for merger plans. If this resistance reflected correct anticipations of future merger effects on service production, the effect on output and quality in schools located in former surrounding municipalities could be negative. The rich individual by school by municipality data available to us, makes it possible to test this hypothesis.

This paper estimates the effect of school district size through municipal mergers using a school fixed effects strategy. Municipality mergers are found to significantly increase student income in adulthood by 2-3%, while the effect on educational attainment is generally positive, but not precisely estimated. To enhance the understanding of possible mechanisms behind this important result, we further investigate possible heterogeneous effects by school location and the effect of mergers on school characteristics and fiscal variables, using the same difference-in-differences approach but with municipalities as the unit of analysis.

Our results clearly show that the income effect is driven by students enrolled in schools in pre-merger municipalities surrounding the former city. The effect on students enrolled in schools located in the pre-merger city is numerically very small and far from significant. Thus, the hypothesis that former surrounding municipalities resisted merger because of correct anticipations of negative future merger effects on service production and quality is not supported by the empirical results. Rather the evidence suggests the opposite. Output and quality as measured by our variables increased in these former surrounding municipalities. The former cities became administrative centers in the new municipalities. The finding is consistent with the hypothesis that students enrolled in schools in former surrounding municipalities took advantage of potential gains in existing administrative quality in the former cities, although further research is needed to confirm this interpretation.

When deciding whether to merge municipalities together, proponents argue that

larger municipalities increase efficiency, while opponents argue that the population is further removed from their elective officials. Results from this paper suggest that municipality mergers can have positive effects on school outputs measured by years of education and income in adulthood, lending support to the proponents of municipality mergers.

Chapter 4: Grade variance

Astrid Marie Jorde Sandsør

What are the effects of the individual distribution of skills on school attainment and school performance? We know that cognitive skills are an important predictor for future outcomes for the individual, including education and labor market outcomes (Murnane, Willett, and Levy, 1995; Herrnstein and Murray, 2010; Heckman, 1995), and aggregate measures of cognitive skills are important for of economic growth and development (Hanushek and Woessmann, 2008; Hanushek and Kimko, 2000). However, for a given average level of skills, is it better that skills are evenly divided between subject areas or is it better to be particularly good at some subject area?

One measure of cognitive skills is student grades received in school, commonly measured as the grade point average. Grades are highly correlated with short-term and long-term outcomes such as educational attainment and income. Additionally, grades have direct consequences for students, by for instance forming part of the college admission decision and determining their post-education job qualifications. Grade point average captures the first moment of the individual grade distribution, the mean. The second moment of the distribution, the variance, is a measure of grade dispersion; how far the grades are from the individual's mean. For a given grade point average, which student might be expected to have higher educational attainment; the student with high or low grade variance?

On the one hand, grades might reflect non-cognitive skills, such as motivation, perseverance and conscientiousness which have been shown to be meaningful predictors of educational, labor market and behavioral outcomes. If high grade variance is associated with low non-cognitive skills and vice versa, then a negative relationship between grade variance and educational attainment is expected. On the other hand, grades might mainly reflect knowledge in the subject, i.e., cognitive skills. As higher education allows students to specialize in their preferred field, high variance students, who are particularly good in some subjects, might be expected to have a higher educational attainment.

As there are reasons to believe that grade variance could be either positively

or negatively associated with educational attainment, this makes grade variance particularly interesting to study empirically. Finding a negative association between grade variance and educational attainment, especially at the lower end of the grading distribution, supports the non-cognitive skills hypothesis while finding a positive association, especially at the upper end of the grading distribution, supports the generalist/specialist hypothesis.

In order to investigate the importance of grade variance empirically, I use three different data sources; The U.S. National Longitudinal Survey of Youth, 1979 (NLSY79), Norwegian register data (NRD) and data from the Character Development in Adolescence Project (CDAP). The NLSY79 is a longitudinal survey with a nationally representative sample of young Americans first interviewed in 1979 and includes high school transcript data, educational attainment and socioeconomic characteristics. The NRD contains the entire population of students graduating from lower secondary education in Norway from 2002-2004 and includes transcript data, educational attainment and socioeconomic characteristics. The CDAP is a longitudinal survey of middle school students and their teachers from 8 different schools and includes transcript data along with various self-reported and teacher-reported measures of non-cognitive skills.

The NLSY79 and NRD are both used to investigate the association between grade variance and educational attainment and whether the association differs across the grading distribution or by gender. The NLSY79 includes long-run educational outcomes while the NRD only includes short-run educational outcomes. In Norway, grades are the main determinant of acceptance into upper secondary and higher education, and grading practices are monitored by central authorities, reducing potential measurement error. Along with the richness of register data, this allows for a more detailed analysis in the NRD than in NLSY79. By investigating data from two different countries, I am able to investigate whether the association between grade variance is context specific or more general.

Next, the paper investigates how grade variance is associated with cognitive and non-cognitive skills. The NLSY79 includes measures of cognitive and non-cognitive skills previously used by Heckman, Stixrud, and Urzua (2006) while a subset of grades is used as measures of cognitive and non-cognitive skills in the NRD. However, in both data sets the measures of cognitive and non-cognitive are simple and may not be capturing the skills that could be expected to be associated with grade variance. The CDAP includes grades together with a rich set of non-cognitive skills measures allowing for a more robust analysis of non-cognitive skills and grade variance.

For both the United States and Norway, grade variance is found to be neg-

atively associated with educational outcomes. In the NLSY79, grade variance is negatively associated with educational attainment. In the NRD, grade variance is negatively associated with (1) starting the academic track in upper secondary, (2) upper secondary grade point average, (3) graduating from the academic track in upper secondary and (4) continuing on to higher education. Estimates are robust to controlling for socioeconomic characteristics and school fixed effects in the NLSY79 and school by cohort fixed effects in the NRD. The estimate for grade variance is negative across the grading distribution for both countries and no significant differences are found between boys and girls.

The association between grade variance and educational outcomes remains negative when including measures of cognitive and non-cognitive skills. In the NLSY79, the estimate for grade variance is reduced when adding cognitive skills but remains unchanged when adding non-cognitive skills. In the NRD, adding cognitive and non-cognitive measures do not change results in a systematic way. The CDAP data confirm that grade variance does not seem to be related to non-cognitive skills. While the negative association between grade variance and educational attainment supports the non-cognitive skills hypothesis, all results are robust to adding measures of non-cognitive skills which does not support this hypothesis. Results support the alternative hypothesis that being a generalist rather than a specialist is beneficial for educational attainment.

If institutions are interested in students with high ability and effort, but only use grade point average in the admission decision, they may not be accepting the best students. Students with low grade variance who are just below the grade point average cutoff are likely to outperform student just above the cutoff with high grade variance. My findings support that institutions should take grade variance, or other measures of skill, into account in admission decisions.

References

Almond, D., and J. Currie (2011): "Human Capital Development before Age Five," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 4, chap. 15, pp. 1315–1486. Elsevier.

Angrist, J., E. Battistin, and D. Vuri (2015): "In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno," IZA Discussion Papers 8959, Institute for the Study of Labor (IZA).

Angrist, J. D., and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114(2), 533–575.

Baker, M. (2011): "Innis Lecture: Universal early childhood interventions: what is the evidence base?," *Canadian Journal of Economics*, 44(4), 1069–1105.

Baker, M., J. Gruber, and K. Milligan (2008): "Universal Child Care, Maternal Labor Supply, and Family Well-Being," *Journal of Political Economy*, 116(4), pp. 709–745.

Barnett, W. S., and C. R. Belfield (2006): "Early Childhood Development and Social Mobility," *Future of Children*, 16(2), 73–98.

Becker, G. S. (1964): *Human Capital*. New York: Columbia University Press.

Becker, G. S. (1978): *The Economic Approach to Human Behavior*. University of Chicago press.

Berry, C. R., and M. R. West (2010): "Growing pains: The school consolidation movement and student outcomes," *Journal of Law, Economics, and Organization*, 26(1), 1–29.

Beuchert, L. V., M. K. Humlum, H. S. Nielsen, and N. Smith (2015): "The Short-Term Effects of School Consolidation on Student Achievement: Evidence of Disruption?," *Available at SSRN 2626712*.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2011): "Too Young to Leave the Nest? The Effects of School Starting Age," *Review of Economics and Statistics*, 93(2), 455–467.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. Ter Weel (2008): "The economics and psychology of personality traits," *Journal of Human Resources*, 43(4), 972–1059.

Carneiro, P., C. Crawford, and A. Goodman (2007): "The impact of early cognitive and non-cognitive skills on later outcomes," Discussion paper, CEE DP 92.

Cascio, E. U. (2009): "Do Investments in Universal Early Education Pay Off? Long-term Effects of Introducing Kindergartens into Public Schools," Working Paper 14951, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126(4), 1593–1660.

Clinedinst, M. E., and D. A. Hawkins (2011): "State of college admission," *Washington, DC: National Association for College Admission Counseling*.

Coleman, J. S., E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld, and R. York (1966): *Equality of Educational Opportunity*.

Cunha, F., and J. Heckman (2007): "The Technology of Skill Formation," *American Economic Review*, 97(2), 31–47.

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

Drange, N., T. Havnes, and A. M. J. Sandsør (2012): "Kindergarten for All: Long Run Effects of a Universal Intervention," IZA Discussion Papers 6986, Institute for the Study of Labor (IZA).

Dynarski, S., J. Hyman, and D. W. Schanzenbach (2013): "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion," *Journal of Policy Analysis and Management*, 32(4), 692–717.

Ehrenberg, R. G., D. J. Brewer, A. Gamoran, and J. D. Willms (2001): "Class size and student achievement," *Psychological Science in the Public Interest*, pp. 1–30.

Falch, T., O. H. Nyhus, and B. Strøm (2014a): "Causal effects of mathematics," *Labour Economics*, 31, 174–187.

——— (2014b): "Performance of Young Adults: The Importance of Different Skills," *CESifo Economic Studies*.

Falch, T., B. Strøm, and A. M. J. Sandsør (2015): "Do smaller classes always improve students' long run outcomes?," Working Paper Series 16415, Department of Economics, Norwegian University of Science and Technology.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013): "Long-Term Effects of Class Size*," *Quarterly Journal of Economics*, 128(1), 249–285.

Gordon, N., and B. Knight (2008): "The effects of school district consolidation on educational cost and quality," *Public Finance Review*, 36(4), 408–430.

Gupta, N. D., and M. Simonsen (2010): "Non-cognitive child outcomes and universal high quality child care," *Journal of Public Economics*, 94(1-2), 30 – 43.

Hanushek, E. A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24(3), 1141–77.

——— (2003): "The Failure of Input-Based Schooling Policies," *Economic Journal*, 113(485), F64–F98.

——— (2006): "School Resources," in *Handbook of the Economics of Education*, ed. by E. Hanushek, and F. Welch, vol. 2, chap. 14, pp. 865–908. Elsevier.

Hanushek, E. A., and D. D. Kimko (2000): "Schooling, labor-force quality, and the growth of nations," *American Economic Review*, pp. 1184–1208.

Hanushek, E. A., and L. Woessmann (2008): "The role of cognitive skills in economic development," *Journal of Economic Literature*, pp. 607–668.

Hanushek, E. A., and L. Woessmann (2011): "The Economics of International Differences in Educational Achievement," in *Handbook of the Economics of Education*, ed. by S. M. Eric A. Hanushek, and L. Woessmann, vol. 3, chap. 2, pp. 89 – 200. Elsevier.

Haraldsvik, M. (2012): "Does performance based school choice affect student achievement?," Discussion paper, Doctoral thesis at NTNU 2012:346.

Havnes, T., and M. Mogstad (2011): "No Child Left Behind. Subsidized Child Care and Children's Long-Run Outcomes," *American Economic Journal: Economic Policy*.

Heckman, J. J. (1995): "Lessons from the bell curve," *Journal of Political Economy*, pp. 1091–1120.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24(3), 411–482.

Herrnstein, R. J., and C. Murray (2010): *Bell curve: Intelligence and class structure in American life*. Simon and Schuster.

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

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–75.

Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014): "Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success," OECD Education Working Papers 110, OECD Publishing.

Knudsen, E. I., J. J. Heckman, J. L. Cameron, and J. P. Shonkoff (2006): "Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce," *Proceedings of the National Academy of Sciences of the United States of America*, 103(27), pp. 10155–10162.

Krueger, A. B., and D. M. Whitmore (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.

Leuven, E., M. Lindahl, H. Oosterbeek, and D. Webbink (2010): "Expanding schooling opportunities for 4-year-olds," *Economics of Education Review*, 29(3), 319–328.

Leuven, E., and S. A. Løkken (2015): "Long term impacts of class size in compulsory schooling.," Mimeo.

Leuven, E., H. Oosterbeek, and M. Rønning (2008): "Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway*," *Scandinavian Journal of Economics*, 110(4), 663–693.

Lindqvist, E., and R. Vestman (2011): "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment," *American Economic Journal: Applied Economics*, 3(1), 101–28.

Mincer, J. (1970): "The distribution of labor incomes: a survey with special reference to the human capital approach," *Journal of Economic Literature*, 8(1), 1–26.

——— (1974): "Age and Experience Profiles of Earnings," in *Schooling, Experience, and Earnings*, pp. 64–82. NBER.

Murnane, R. J., J. B. Willett, and F. Levy (1995): "The growing importance of cognitive skills in wage determination," *Review of Economics and Statistics*, 77(2), 251–266.

Norwegian Ministry of Local Government and Labor (1986): *NOU 1986:7: Forslag til endringer i kommuneinndelingen for byområdene Horten, Tønsberg og Larvik i Vestfold fylke (Suggestions to changes in the municipality structure for the city areas Horten, Tønsberg and Larvik in Vestfold county).*

——— (1989): *NOU 1989:16: Kommuneinndelingen for byområdene Sarpsborg, Fredrikstad, Arendal, Hamar og Hammerfest (Municipality structure for the city areas Sarpsborg, Fredrikstad, Arendal, Hamar and Hammerfest).*

Oates, W. E. (1972): *Fiscal Federalism*. New York: Harcourt Brace.

Reingewertz, Y. (2012): "Do municipal amalgamations work? Evidence from municipalities in Israel," *Journal of Urban Economics*, 72(2), 240–251.

Ruhm, C., and J. Waldfogel (2012): "Long-term effects of early childhood care and education," *Nordic Economic Policy Review*, 1(1), 23–51.

Schultz, T. W. (1961): "Investment in human capital," *American Economic Review*, pp. 1–17.

Shonkoff, J. P., D. Phillips, and N. R. Council (2000): *From neurons to neighborhoods : the science of early child development*. National Academy Press, 1 edn.

Webbink, H. D. (2005): "Causal effects in education," *Journal of Economic Surveys*, 19(4), 535–560.

Wössmann, L., and M. West (2006): "Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS," *European Economic Review*, 50(3), 695–736.

Chapter 1:

Kindergarten for all:
Long-run effects of a universal intervention

Nina Drange, Tarjei Havnes and Astrid Marie Jorde Sandsør

Kindergarten for all: long-run effects of a universal intervention*

Nina Drange† Tarjei Havnes‡ Astrid M. J. Sandsør§

Abstract

Theory and evidence point towards particularly positive effects of high-quality child care for disadvantaged children. At the same time, disadvantaged families often sort out of existing programs. To counter differences in learning outcomes between children from different socioeconomic backgrounds, governments are pushing for universal child care. However, it is unclear how effective programs with universal participation may be at addressing the needs of disadvantaged children. We provide evidence on the long-run effect on schooling of mandating kindergarten at age 5–6. Our identifying variation comes from a reform that lowered school starting-age from 7 to 6 in Norway in 1997. The new program was designed as a low intensity kindergarten program, similar to voluntary child care programs available before mandating. Our precise DD estimates reveal hardly any effect, both overall, across subsamples, and over the grading distribution. A battery of specification checks support our empirical strategy.

Keywords: kindergarten, early childhood intervention, distributional effects, difference-in-differences, child care, child development

JEL codes: J13, H40, I28

*Thanks to Michael Baker, Torbjørn Hægeland, Magne Mogstad, Mari Rege, Marianne Simonsen and Kjetil Telle, as well as participants at a number of seminars and conferences. The project is part of the research activities at the ESOP center at the Department of Economics, University of Oslo. ESOP is supported by The Research Council of Norway. Havnes also gratefully acknowledges support from Statistics Norway and funding from The Research Council of Norway (S/194339 and S/212305). Drange also gratefully acknowledges support from The Research Council of Norway (160965/V10).

†Statistics Norway

‡Department of Economics, University of Oslo

§Department of Economics, University of Oslo

1 Introduction

Universally available child care of high quality can benefit child development, also in the long run (Almond and Currie, 2011). Returns are often found to be particularly high for children from disadvantaged families.[1] At the same time, children from disadvantaged families are underrepresented in existing programs. This sorting into the programs coupled with particularly large estimated benefits among disadvantaged children, suggests a potentially strong social gradient in expanding or mandating early childhood interventions (Barnett and Belfield, 2006). Indeed, in an effort to counter differences at school entry depending on social background, many countries are currently moving towards subsidized child care available for the general population.[2]

Policies and proposals promoting universal interventions in early childhood pose a challenge to the existing literature, which has reserved most of its attention for programs targeted at disadvantaged children. Existing studies on universally available programs typically reveal the impact on children from families with a strong preference for out-of-home care. For instance, Baker, Gruber, and Milligan (2008) and Havnes and Mogstad (2011) study the introduction of a universally available program but with actual enrollment being far from universal, while Gupta and Simonsen (2010) explicitly exploit rationing of child care for identification. Since both theory and evidence point towards important heterogeneity in the effects of early childhood interventions, it remains an open question how well the current evidence can inform about the impact of truly universal interventions. In particular, it is unclear how effective programs with universal participation may be at addressing the needs of disadvantaged children.

In the current paper, we provide evidence on the long-run effect on schooling of a reform that mandated kindergarten at age 5–6. We first consider the impact on children's school performance at the end of compulsory schooling at age 15–16. We also consider the impact on high school dropout (age 18) and on enrollment in the academic track (age 16) in upper secondary school. These are interesting in

[1] Havnes and Mogstad (2012) document large heterogeneity in the effects on adult outcomes from child care for 3–6 year old children in the late 1970s in Norway. Ludwig and Miller (2007) interpret the effects of the U.S. Head Start as an upper bound because children are among the most disadvantaged. Further, effects found in the targeted Perry Preschool project (e.g. Karoly, Kilburn, and Cannon, 2005) are larger than what could plausibly be expected in the general population.

[2] For instance, U.S. President Obama stated in his 2013 State of the Union Address that he wants to "make high-quality preschool available to every child in America". In Europe, the European Union Commission proclaims that early childhood education and care (ECEC) "is the essential foundation for successful lifelong learning, social integration, personal development and later employability" (European Union, 2011, p. 1).

their own right, and help confront the concern of fading out of cognitive effects from early intervention programs, even when long-term effects on substantive outcomes may persist.[3] Our identifying variation comes from a 1997-reform in Norway that lowered school starting age from seven to six. The new program for six year olds was designed as a low intensity kindergarten program, aimed at preparing children for school by learning through play, similar to early U.S. kindergarten programs (Cascio, 2009). The goal of the new program was to counter differences in learning outcomes between children from different socioeconomic backgrounds. While disadvantaged children were thought to benefit most from kindergarten programs, they were strongly underrepresented in the existing voluntary programs prior to the reform.

Because the implementation of the reform was nationwide, the most direct assessment compares cohorts just young enough to be affected with cohorts just old enough not to be affected. An immediate objection to this strategy is that we may be confounding effects of the policy with unrelated cohort effects. To get around this issue, we take advantage of voluntary enrollment in child care prior to the reform. Since the new program for six year olds bears strong resemblance to kindergarten programs that were widely available prior to the reform, it should have little impact on children that would voluntarily enroll in such programs. This motivates a difference-in-differences (DD) approach where we compare outcomes before and after the implementation of the mandatory kindergarten reform, of children who enroll in voluntary kindergarten at age six (i.e. the control group) and children who do not enroll in voluntary kindergarten at age six (i.e. the treatment group). Because voluntary enrollment at age six is unobserved by definition after the reform, we use enrollment in child care at age five to determine treatment. This should be a good proxy since children who are enrolled at age five are almost universally enrolled at age six.

Results reveal that the program had little impact on affected children. In our baseline estimation, the precisely estimated effect on the child's school performance is negative but below 2 % of a standard deviation. Meanwhile, we find a modest increase in high school dropout rates, and no impact on academic tracking in upper secondary school. These results are robust to including or excluding a large set of observable characteristics, as well as a battery of specification checks confronting the key identifying assumption of common trends in treatment and comparison groups. Importantly, we find no evidence of a separate effect of the reform on

[3]See for example Heckman, Moon, Pinto, Savelyev, and Yavitz (2010) or Heckman, Pinto, and Savelyev (2013).

our comparison group that may attenuate effects on the treatment group, nor of a delayed effect of the reform on later cohorts. We also find no evidence of important heterogeneity when we look across subsamples reflecting the child's background and home environment, across different segments of the grading distribution, or across school subjects where we may expect children to benefit from different types of skills.

To help interpret our estimates, we take a close look at the contents of the program, which was specifically intended to be play-oriented, with little focus on specific learning activites. As a comparison, the program appears to be quite similar in content to the early U.S. kindergarten programs, as its focus was more on children's social development than on academic training, though the compulsory nature of the Norwegian program is an important difference.[4] The program also seems comparable to the U.S. Head Start program, with its low intensity educational content, as well as similar costs and contents.[5] While the program we study served the entire population of 5–6 year olds, however, Head Start serves children 3–5 years old and is targeted at poor families.

Our paper contributes to the rapidly increasing literature on how early childhood interventions in general, and kindergarten programs in particular, can promote the formation of skills in children.[6] This literature is divided into two distinct branches, one focussed on targeted programs, the other focussed on universal programs available to the general population. While studies of targeted programs often find positive effects,[7] the literature on universal programs is smaller and findings are mixed.[8]

[4]See Cascio (2009) for details and discussion on development of the U.S. kindergarten program, and Norwegian Ministry of Education (2010) on the Norwegian program.

[5]See Deming (2009) for details and discussion on Head Start.

[6]For recent reviews, see Almond and Currie (2011), Ruhm and Waldfogel (2012), or Baker (2011). Our paper also relates to the literature on early enrollment into formal schooling (see e.g. Leuven, Lindahl, Oosterbeek, and Webbink (2010) or Black, Devereux, and Salvanes (2011) for an overview). An important issue in this literature has been to resolve the collinearity of age at test and age at school start. This is not an issue in our case, since age at test is both common across treatment groups and unaffected by the reform. However, the literature on child care and early childhood interventions may in general be said to face a similar collinearity between age at program start and years of enrollment. As in the rest of the literature, we estimate the combined effect of an additional year in kindergarten and lower age of entry.

[7]The Perry Preschool and Abecedarian programs are examples of targeted randomized programs (see Barnett (1995) and Karoly, Kilburn, and Cannon (2005) for surveys of the literature), while the U.S. Head Start program provides an example of a targeted non-randomized program (see e.g. Currie (2001) or McKey, Condelli, Ganson, Barrett, McConkey, and Plantz (1985) for a review of the findings).

[8]Several studies from Canada show a negative impact on a variety of child outcomes (Baker, Gruber, and Milligan, 2008; Lefebvre and Merrigan, 2008b; DeCicca and Smith, 2013), while Cascio (2009) and Gupta and Simonsen (2010) find essentially no impact from child care programs in the United States and Denmark, respectively. In contrast, positive impacts on long-run outcomes are found from child care programs in several countries, including the United States (Fitzpatrick, 2008), Uruguay (Berlinski, Galiani, and Manacorda, 2008), Norway (Havnes and Mogstad, 2011), Germany (Dustmann, Raute, and Schonberg, 2013; Felfe and Lalive, 2013), and Spain (Felfe,

Perhaps as a consequence, the discussion on child care policies is based largely on the targeted literature and descriptive evidence, even when the policies considered are universal.

There are several reasons why effects from programs targeted at disadvantaged children could differ importantly from more universal programs, as discussed by Baker (2011). First, the effect of such programs is related to the alternative mode of care had the programs not been in place. Since disadvantaged children would be expected to have poorer alternatives, they likely have more to gain from interventions (Knudsen, Heckman, Cameron, and Shonkoff, 2006). Second, targeted interventions are often quite intensive, sometimes including home visits, nutritional advice and several years of daily activities. In comparison, a program serving a large part of the population will necessarily have to provide a less intensive intervention. This might produce effects from large-scale and universal programs that differ substantially from the effects of intensive small-scale interventions.

The current paper contributes to the literature on universal child care programs in two distinct ways. First, since kindergarten is not rationed prior to mandating, the estimated effect should derive from the particular group of children that do not voluntarily enroll. Our study therefore provides a rare opportunity to learn about the group of never-takers, to use the terminology of Imbens and Angrist (1994). This is of particular interest since these families may have quite different characteristics compared to families that select into child care voluntarily, many of which may be unobserved. If so, then existing estimates may tell us little about the potential effect of child care among these children. Second, while the program we study is universal, the reform may be viewed as targeted since affected children come disproportionately from disadvantaged families. Our results may therefore shed light on how a universal low intensity program can improve outcomes among the disadvantaged. That is, can the positive effects for the disadvantaged, often seen from targeted interventions, be reproduced in a universal program? Our evidence suggests that this is not the case. This is true, even though the estimates likely reflect shifts mostly from parental care, rather than informal care.

Our results differ from some previous studies that have found positive effects of child care, particularly for low income children. One reason may be that the larger operations and broader scope involved in a universal compared to a targeted program may come with particular challenges, for instance by making it harder

Nollenberger, and Rodríguez-Planas, 2012). Also, while the picture is somewhat mixed, the most robust evidence on the U.S. Head Start program tends to show positive effects on long-run outcomes such as high school dropout, college attendance and crime (Currie and Thomas, 1995; Garces, Thomas, and Currie, 2000; Ludwig and Miller, 2007; Deming, 2009).

to see children's needs and to tailor activities to these needs. This might suggest that the unstructured child-centered approach to instruction (Stipek, Feiler, Daniels, and Milburn, 1995), which has been a hallmark of low intensity child care programs, may be less suitable in a universal program. An alternative interpretation is that parents who chose not to send their children to early education when such a program was available and affordable, may have done so partly because they expected little benefits to their children. In any case, while early childhood investments through subsidized child care can be an important tool in facilitating equal opportunities, our evidence emphasizes that this is hardly automatic, and that the structuring of the program and its content could be key to generating the intended benefits.

The paper proceeds as follows. We first discuss the institutional background for the 1997-reform in Section 2. Section 3 describes our data and gives descriptive statistics while Section 4 discusses our empirical approach. Section 5 then presents our main results, before Section 6 presents a battery of specification checks and investigates potential mechanisms. Section 7 concludes.

2 Background

Until 1997, Norwegian children started school in August the year they turned seven. This was late compared to children in most western countries.[9] At the same time, slots in child care institutions were widely available following a child care reform in 1975. In 1996, 89 % of non-immigrant families enrolled their six year olds in a kindergarten program.[10] However, from the mid-1980s, there was widespread worry that children entered school on different footings, depending on their socioeconomic background.

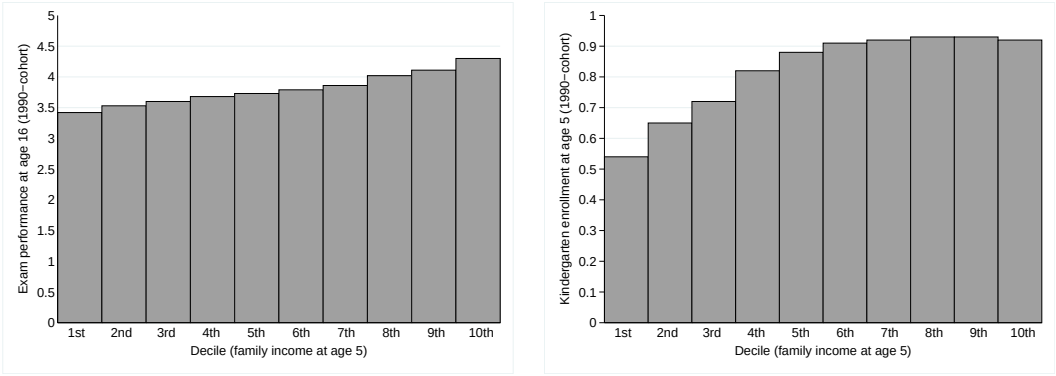
Figure 1 shows the strong social gradient in school performance and kindergarten enrollment.[11] In Panel (a), we draw the average grade of students at exams administered at the end of compulsory school, in the deciles of family income at age five. The figure shows a strong positive relationship between the two. On average, children in the lowest decile, with family income of about USD 16,000, receive a grade of less than 3.5, while children in the upper decile, with family income of about USD 170,000, receive a grade of almost 4.5.[12] This difference in exam performance is equivalent to a difference of just under one standard deviation, comparable to the

[9]For instance, school starting age in Germany, France and the United States was six, while England had a starting age of five.

[10]For simplicity, we use age a to refer to the year the child turns a years old in the following.

[11]For details, see Table A5 in Appendix A.

[12]Throughout, we refer to 2011-USD adjusted using the consumer price index, USD/NOK = 6.



(a) School performance (b) Enrollment in kindergarten

Figure 1: Social gradient in school performance and enrollment in kindergarten among children born in 1990.

Note: Family income is measured in 1996, when the child is five years old, adjusted for CPI-growth, and converted to USD using $USD/NOK = 6$. Average exam grade and enrollment in kindergarten refers to the mean among children from families with income in each decile of the distribution of family income. Data descriptions and variable definitions are found in Section 3.

90-10 income achievement gaps in the United States reported by Reardon (2011).

At the same time, children in lower deciles have a much lower probability of being enrolled in kindergarten. Panel (b) of Figure 1 shows that enrollment in kindergarten at age five among children in the lowest decile of family income is just over 50 %, compared to over 90 % for deciles 6–10. This serves on the one hand to illustrate the political background for the reform discussed above, and on the other hand to show that the children that were affected by the reform should come disproportionately from low-income familes.

While children enrolled in formal child care were offered school preparation in kindergarten groups within their child care center, this was not available for children not enrolled in formal child care.[13] On this background, a proposal to lower the mandatory school starting age from seven to six was widely discussed. Compulsory programs for six year olds, as opposed to voluntary kindergarten programs, would expose all children to the same educational program, and was argued to counter differences in learning outcomes between children from different socioeconomic backgrounds. A reform was finally proposed in a government White Paper published in the spring of 1993 (Norwegian Ministry of Education, 1992-93), and passed the Norwegian Parliament in May 1994 (Norwegian Ministry of Education,

[13]Voluntary programs for six year olds were allowed on school grounds from 1991, managed by kindergarten teachers (Norwegian Ministry of Education, 1990-91, Ot.prp. nr. 57). Government support was the same as for six year olds enrolled in kindergarten, and the parental copayment and educational content of the program were essentially the same as that offered in regular child care institutions. In the remainder, we do not distinguish between the two, as we cannot identify in which program a particular child was enrolled.

1993-94). The reform was implemented in August 1997, at the start of the 1997–1998 school year. The first children affected were those born in 1991, who started school in August 1997, the year they turned six years old.

Note that the cutoff for school starting age in Norway is January 1st. In Norway schools employ strict enrollment rules, and nearly all children start school the year they turn the school starting age. Any exemption from this rule requires a formal application from the parents which then has to be approved by specialists and decided upon by the local government.

Structural content. The group size was capped at 20 children, under supervision of two kindergarten teachers, identical to the previous child care programs. The government cost of the program was about 8,800 USD per child per year, similar to the cost of the previous child care programs which cost about 9,700 USD per child per year. There was no parental copayment for the core four hour program, but the voluntary after-school program required a copayment of about 170 USD per month on average. This was a reduction from the copayment for center-based kindergarten, which ranged from about 290 to 630 USD per month, depending on income. The reduced cost could imply that the reform caused an economic windfall for children who would attend kindergarten in the absence of the reform, with possibly positive effects on their school performance. We investigate the potential for income effects extensively below, finding little cause for concern.

Educational content. The new program was aimed at combining the best of school and kindergarten traditions. These were grounded in the tradition of social pedagogy that dominates child care practices in Norway since the 1970s.[14] Learning through play was stated as essential, and formal learning was given little credence (Norwegian Ministry of Education, 1992-93). The curriculum specifically stated that *"The first year is to have a distinct kindergarten character, and one has to emphasize learning through play and age-mixed activities throughout elementary school [years 1–4]"* (Norwegian Ministry of Education, 1996a).

Note that the curriculum for all grades in primary school was revised, and implemented for grades 1, 2, 5 and 8 in 1997, grades 3, 6 and 9 in 1998 and the remaining grades in 1999 (Norwegian Ministry of Education, 1996b). Children in our main sample, born 1990–1991, were therefore subject to the same, new curriculum throughout primary school.

[14]The social pedagogy tradition for early education has been especially influential in the Nordic countries and Central-Europe. In contrast, a so-called pre-primary pedagogical approach to early education has dominated many English and French-speaking countries, favoring formal learning processes to meet explicit standards for what children should know and be able to do before they start school.

In the new mandatory kindergarten program, the minimum requirement was one teacher or kindergarten teacher for every 18 children. By comparison, the minimum requirement for pedagogical staff in child care centers for six year olds was one per 14–18 children. Beyond this, the municipalities should themselves judge the need for further staff, but they had to secure sufficient care for the children (Norwegian Ministry of Child and Family Affairs, 1995).[15] Similarly, in the new kindergarten program integrated into schools, in addition to the kindergarten teacher, assistants were hired depending on the size of the group. Starting in 1991, both kindergarten teachers and school teachers were allowed to work with six year olds. This was part of the gradual implementation of voluntary programs for six year olds on school grounds. After the 1997-reform was passed in parliament, both kindergarten teachers and teachers could work in first grade, while some continued education was needed for kindergarten teachers to work in grades 2–4. The reform explicitly aimed to bring together the best of school and kindergarten traditions, and bringing kindergarten teachers into the first grades of school was part of this goal. The transfer of kindergarten teachers from child care centers into schools, meant that six year olds in 1997 were likely to experience a similar pedagogical environment to the one they would have experienced in absence of the reform.

The teaching requirement for the new first graders was set to 20 hours per week (Norwegian Ministry of Education, 1992-93). To ensure the care of six year olds during normal work hours but outside school hours, the government also expanded the access to the pre-existing after-school program, available for children in grades 1–4. After-school programs were available throughout our period of study, were subject to similar requirements as regular child care providers, and were usually situated on school premises. The programs offered free play under the supervision of non-qualified adults, with no educational content.

Other reforms. We may worry that there were other reforms that could have affected our cohorts differently. However, the closest reform in primary education prior to the 1997-reform was implemented in 1986, while there were no additional reforms until the start of the school year 2007–2008. This ensures that the 1990 and 1991 cohorts completed their entire compulsory schooling with the national curriculum introduced in 1997. A nationwide cash-for-care reform was implemented in 1998,

[15]The head kindergarten teacher was responsible for planning, observing, collaborating and evaluating the work being done, under the requirements specified in the regulations for subsidized child care. Teachers typically worked closely with one or two assistants, and were responsible for the educational programs in separate groups of 6–18 children and for day-to-day interaction with parents. The kindergarten teacher education is a college degree, while there are no educational requirements for assistants.

and expanded in 1999, paying families with children below two years old (from 1998) and three years old (from 1999) that did not utilize subsidized child care a substantial monthly cash allowance.[16] While this reform did not affect the children in our sample directly, it could have had an effect on younger siblings and therefore an indirect effect on the children in our sample (Bettinger, Hægeland, and Rege, 2014). However, the impact of the cash-for-care reform does not differ between children born in 1990 and 1991, which constitute our baseline estimation sample.[17] This suggests that the cash-for-care reform does not pose a threat to our empirical strategy.

3 Data

Dataset and variables. Our data are based on administrative registers from Statistics Norway. Specifically, we use a rich longitudinal database which covers every resident from 1992 to 2007. It contains individual demographic information (e.g. sex, age, immigrant status, marital status, number of children), socioeconomic data (e.g. years of education, income, employment status), and geographic identifiers for municipality of residence. Information on school performance, educational attainment and school enrollment for every individual is based on annual reports from Norwegian educational establishments. Income and employment data are collected from tax records and other administrative registers. Household information is from the Central Population Register, which is updated annually by the local population registries and verified by the Norwegian Tax Authority. We also have access to registry data on municipal child care coverage reported by the child care institutions themselves. The reliability of Norwegian register data is considered to be very good, as documented by highest ratings received in a data quality assessment prepared for the OECD by Atkinson, Rainwater, and Smeeding (1995).

Estimation sample. We start with the universe of children born 1990–1991, who reside in Norway the year they turn five years old and who graduate from lower secondary school in 2005–2006. We then restrict our sample to children born to native-born parents, constituting about 96 % of the population, in order both to focus our study on the effect of mandating, and to sidestep problems of comparability

[16] See Schøne (2004) or Drange and Rege (2013) for a detailed description of the cash-for-care reform.

[17] To investigate this directly, we estimated our baseline DD model in equation (1) using as dependent variable a dummy equal to one if the child has a younger sibling born 1996 or later (i.e. partly or fully eligible for the subsidy) and zero otherwise. The estimate is almost exactly zero (0.004, $SE = 0.007$).

between native and immigrant children. Our paper does not, therefore, speak to the debate on early interventions to provide language training among non-native speakers. We also exclude a handful of children with missing values on our dependent variable. Rather than exclude children with missing values on control variables, we construct dummy variables for missing and include these in our regressions. Our main sample then consists of 111,397 individuals, of which just over 16 % are in the treatment group. In our extended sample, we consider the analogous population of children born 1988–1992.

Measuring kindergarten enrollment. There is, unfortunately, no register of individual kindergarten or child care enrollment. However, parents may claim the cost of child care as a deduction on their earned income. To identify whether a child is enrolled in kindergarten, we therefore follow Black, Devereux, Løken, and Salvanes (2014) in using a binary variable equal to one if the child's parents claimed a tax deduction for child care for the year the child turned five years old. Of course, if a child has siblings, we cannot verify which of the children the deduction is claimed for (if not all). To get around this issue, we assume that child care enrollment is monotonous in age, such that older children are in child care whenever younger children are in child care. This ensures that at least the older child in child care age is enrolled whenever the parents claim the deduction. We might worry that low income households did not take the tax deduction even if eligible. However, payment for child care for each child is often reported directly from the child care provider to the tax authorities. Indeed, since 1994 all public child care institutions have been required to do so.[18] To verify that our measure of kindergarten enrollment is sound, we have calculated the municipal enrollment implied by this measure. We then compare these numbers to the actual enrollment from administrative registers, reported by the child care institutions themselves. The correspondence is very high, with a correlation of about 0.94 (see also Figure A1 in the appendix).

Measuring school performance. Our main outcome is an average of grades on nationally administered end-of-school exams. At graduation from compulsory school, students are tested on two or three exams in randomly drawn theoretical subjects—one or two written exams and one oral exam. The written exam is uniform across the country and provided by the Central Education Authority, and is corrected by external evaluators who typically grade exams from several schools simultaneously. The oral exam is also evaluated by an external examiner, and takes place at the

[18] While the vast majority of deductions claimed are for child care costs, some other costs may also be claimed under the same statute, e.g. outlays for support of children with disabilities or with other special needs.

school at which the student is enrolled. Grades are awarded on a scale from one to six, where six indicates excellence and one indicates very little competence (in our estimations, we standardize grades to mean zero and standard deviation one). Grade retention is illegal, hence all students are allowed to graduate regardless of their grades. In addition, teachers assign each student grades in 12–13 subjects, based on performance throughout the year. There are nine theoretical subjects and four practical subjects.[19]

Measuring high school dropout. In our data we can observe whether an individual is enrolled in education and how many years they have successfully completed. We define high school dropout as either not being enrolled in education, or not being on year for age in graduating year. That is, we code high school dropout as not being registered in your 13th year of education in the fall of the year you turn 18.[20] Note that this definition is somewhat strict, since it requires that students are not delayed.

Measuring academic track. In Norway, students are first tracked when they start upper secondary school. There are two main tracks (which are divided into 13 more specialized sub-disciplines): The academic track which is required for entry into university and college studies, and the vocational track which qualifies for a practical occupation. To consider whether the reform had an impact on academic tracking, we use a dummy equal to one if the child started on the academic track in the year following graduation from compulsory schooling, i.e. at age 16. Note that enrollment in upper secondary school is almost complete in these cohorts, with about 94 % of students enrolling in one of the two tracks. We have also estimated the effect on the decision to enroll, finding no impact of the reform. If a student does not enroll in upper secondary school the year following graduation, he or she is excluded from these estimations.

Covariates. To account for possible observable changes in composition between years, we include a number of child and parent characteristics in our analysis, measured when the child is five years old. Child characteristics include municipality of residence, gender, number of siblings, and finally a dummy measuring if the child lived in a densely populated area. Background characteristics include a dummy measuring if the mother/father worked full time, a dummy for whether the mother/father

[19]Theoretical subjects are written and oral Norwegian, written and oral English, mathematics, nature and science, social science, and religion. Practical subjects are home economics, physical education, music, and arts and crafts.

[20]Final graduation from high school should occur the year they turn 19 in the academic track and the year they turn 20 in the vocational track. This information is not yet available in the data.

completed high school and a dummy indicating if the mother/father finished a college education. In addition, we include a dummy capturing missing observations on mothers/fathers education. Further, we include a dummy that captures whether the mother/father was younger than 22 when the child was born. We also include a dummy for having missing observations on either the mother or the father. If both parents are missing we exclude the observation. Finally, we include a dummy capturing if one or both parents received welfare benefits, a dummy measuring if the family was low income (defined as earnings below the 10th decile in the family income distribution in the cohort born in 1990), and a dummy capturing if the child lives with only one of its parents.

Descriptive statistics. Means of the outcome variables are presented in Figure 2 of Section 4, and are discussed there. In Table 1, we present characteristics for the entire sample in the first two columns, and differences between the two groups by cohort in the remaining columns. All covariates are measured when the child is five years old. We see no evidence of changes over time for characteristics of children or their parents between the treated and the comparison group. As discussed above, it is clear, however, that the treated children to a greater extent come from families with younger and less educated parents, and are more likely to belong to a family on welfare and/or to a single parent family. They are also overrepresented in the low income family group. This suggests that the children in our treatment group have a more disadvantaged background, in line with the expressed motivation of the policy-maker (Norwegian Ministry of Education, 1992-93).

4 Empirical strategy

Because the implementation of the reform was nationwide, the most direct assessment of how it affected children's long-term outcomes compares cohorts just young enough to be affected with cohorts just old enough not to be affected. An immediate objection to this strategy is that we may be confounding effects of the policy with unrelated cohort effects. To get around this issue, we exploit the temporal and spatial variation in pre-reform kindergarten enrollment in a difference-in-differences setup. Ideally, we want to compare the child outcomes before and after the implementation of the mandatory kindergarten reform of children who would enroll in voluntary kindergarten at age six (i.e. the control group) and children who would not enroll in voluntary kindergarten at age six (i.e. the treatment group). Our basic

Table 1: Summary statistics

	Mean (SD)		**T − C, by cohort**				
	Treated (T)	*Comp. (C)*	*1988*	*1989*	*1990*	*1991*	*1992*
A. Child and family characteristics							
Female	0.49 (0.50)	0.49 (0.50)	0.00	0.00	0.00	0.01	0.01
1 sibling	0.40 (0.49)	0.51 (0.50)	-0.11	-0.10	-0.11	-0.12	-0.12
2 siblings	0.28 (0.45)	0.24 (0.43)	0.04	0.04	0.04	0.05	0.02
3 siblings +	0.12 (0.32)	0.06 (0.23)	0.05	0.06	0.06	0.07	0.06
Densely pop. area	0.57 (0.49)	0.58 (0.49)	0.00	0.00	0.00	0.00	0.00
On welfare	0.19 (0.40)	0.08 (0.26)	0.13	0.13	0.12	0.14	0.16
Low income	0.08 (0.27)	0.00 (0.05)	0.08	0.08	0.07	0.08	0.09
Single parent	0.29 (0.46)	0.17 (0.38)	0.12	0.13	0.12	0.11	0.14
B. Mother characteristics							
Employed	0.18 (0.38)	0.70 (0.46)	-0.55	-0.54	-0.52	-0.53	-0.64
– full time	0.07 (0.26)	0.33 (0.47)	-0.26	-0.27	-0.26	-0.28	-0.32
High school	0.34 (0.47)	0.54 (0.50)	-0.21	-0.22	-0.20	-0.22	-0.26
College	0.13 (0.34)	0.28 (0.45)	-0.17	-0.16	-0.15	-0.16	-0.20
Young mother	0.21 (0.41)	0.13 (0.33)	0.10	0.10	0.08	0.07	0.11
C. Father characteristics							
Employed	0.63 (0.48)	0.73 (0.44)	-0.10	-0.10	-0.10	-0.10	-0.11
High school	0.49 (0.50)	0.60 (0.49)	-0.13	-0.12	-0.11	-0.10	-0.12
College	0.18 (0.39)	0.27 (0.44)	-0.10	-0.09	-0.09	-0.08	-0.10
Young father	0.07 (0.26)	0.05 (0.21)	0.04	0.04	0.03	0.03	0.05

Note: The treatment group includes children whose parents did not report a tax deduction for child care expenses the year the child turned five. Outcome and control variables are defined in Section 3. Standard deviations are in parentheses.

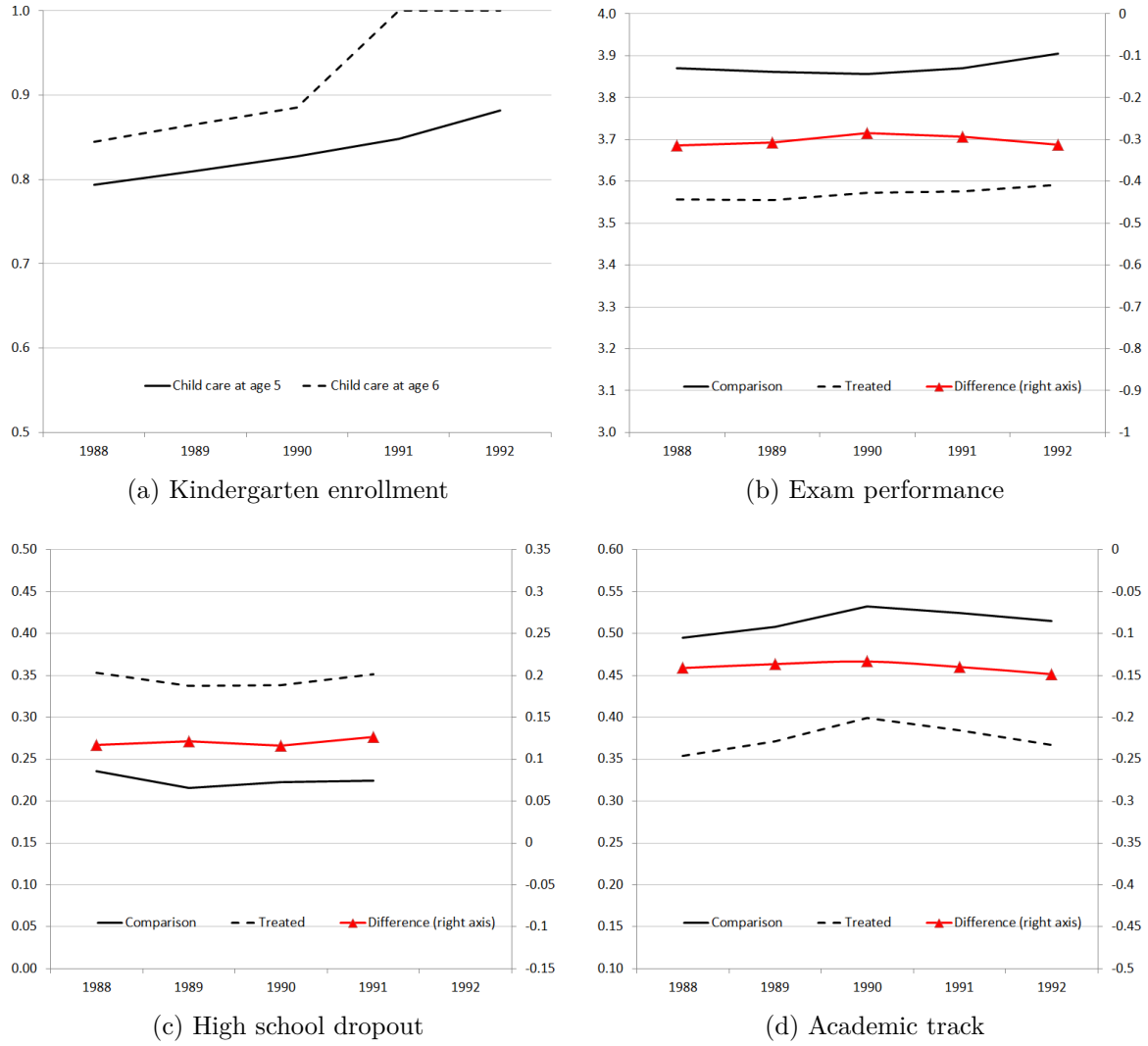
difference-in-differences (DD) model estimated by OLS, can then be expressed as

$$Y_{it} = \alpha_t + \gamma_1 Treated_i + \lambda Treated_i \times Post_t + X_{it}^{'}\beta + \epsilon_{it} \quad (1)$$

where i indexes child, t indexes cohort, $Post_t$ is a dummy equal to one if the child is affected by the reform (i.e. $t \geq 1991$) and zero otherwise, and $Treated_i$ is a dummy equal to one if the child is in the treatment group. Note that the cohort-specific constant term consumes the separate effect of the $Post$-dummy. We estimate the model with and without a large set of control variables for child and parental characteristics X_{it}, including the child's sex, the mother's and the father's age, years of education, and family size (see also Section 3). We also include municipality fixed-effects to capture time-invariant unobserved differences between children from different municipalities. All control variables are measured prior to the impact of the reform and standard errors are robust to heteroskedasticity.

In practice, whether a child would enroll in voluntary kindergarten cannot be observed for post-reform cohorts, since all children are enrolled at age six. To estimate equation (1), we therefore use enrollment in kindergarten at age five to determine treatment. This should be a good proxy since most children who are enrolled in kindergarten at age five are also enrolled at age six. That is, children who are enrolled in child care the year they turn five are placed in the control group, while children who are not enrolled in child care at age five are placed in the treatment group.

Panel (a) of Figure 2 displays the trend in kindergarten enrollment at age five and six in our estimation sample. We note the close relationship between the two series over time in the pre-reform period, where the two lines are virtually parallel. This suggests that enrollment at age five captures the counterfactual evolution of enrollment at age six well. Furthermore, we note that there is no spike in the enrollment of five year olds following the reform, when children age six are no longer taking up places in child care centers. This suggests that there was no discernible rationing of kindergarten for these age groups in our period of study. Finally, we note that kindergarten enrollment in pre-reform years is around 84 %, giving a treatment group of about 16 % of the total sample.



(a) Kindergarten enrollment

(b) Exam performance

(c) High school dropout

(d) Academic track

Figure 2: Kindergarten enrollment and children's schooling outcomes by treatment for cohorts born 1988–1992.

Note: Vertical axes are scaled to approximately one standard deviation. High school dropout is not yet available for the 1992-cohort. Variables are defined in Section 3.

The validity of our DD strategy hinges on the assumption that the trend in school performance among children in the treatment group would have been the same as for children in the control group, in the absence of the reform. As emphasized by Besley and Case (2000), this essentially assumes common time effects and no compositional changes between the treatment and control group. The richness of our registry data allows us to condition on a large set of observable characteristics, to investigate how changes in the composition of the groups may affect our estimates.

To investigate the time effects, Panels (b)–(d) of Figure 2 display mean outcomes of cohorts born 1988–1992 separate for the treatment and control group, and the difference between the two groups over time (on the right axis). The vertical axes are scaled to about one standard deviation in all the figures. The trends are quite flat

and strikingly similar across the treatment and control group throughout the period. The similarity of the trend in the pre-reform period, supports the assumption of common time effects. That there is no jump in the treatment group from the 1991-cohort onwards, nor a divergence in the trends in the post-reform period, is first evidence that the reform had little impact on children's school performance.

One immediate objection to our empirical approach could be that the increasing trend in kindergarten enrollment generates a change in the composition of our treatment and comparison groups. We did not, however, see any evidence of a change in the composition on the observable characteristics of Table 1. Also, the increasing trend in kindergarten enrollment is constant before and after the reform. However, Figure 2 does not reveal any divergence in the school outcomes of the two groups in the pre-reform period, which would be expected if changes in the composition were associated with unobservable determinants of school performance. In Section 6, we also formally challenge our strategy both with a placebo reform in the pre-reform period and by including treatment-specific trends in the specification, neither of which give cause for concern.

Though Section 2 suggests little change in the contents, we could also worry that the new kindergarten program integrated in schools in fact was different from the former program, and thus could have had an effect also on children in the comparison group. We pay close attention to this in our robustness analysis provided in Section 6, finding no support for an effect on the comparison group. To further challenge the validity of our empirical specification, Section 6 also reports results from a series of specification checks.

The difference in enrollment between five and six year olds should not be a threat to the internal validity of our estimates. Higher enrollment at age six than at age five may, however, dilute the estimated treatment effect by misplacing some children in the treatment group who enroll in kindergarten only at age six. Our estimates may therefore be interpreted similar to intention-to-treat (ITT) estimates, and should be scaled in order to arrive at the average treatment effect on the treated. In 1990, 48 % of children who are not enrolled at age five are enrolled at age six, suggesting that only 52 % of the treatment group are in fact affected by the reform. In interpreting our results, we should bear this in mind, scaling the estimated effects by a factor of $1/.52 = 1.9$ to arrive at the average effect on the treated (ATT).[21]

[21]The opposite misclassification is almost completely absent: More than 97 % of children who are enrolled in kindergarten at age 5 are also enrolled at age 6.

5 Empirical results

In this section, we first report estimated mean effects of mandating kindergarten on children's long-term schooling, before we investigate potential heterogeneity in the effects across subsamples and across the grading distribution. All specifications are estimated with municipality fixed effects to account for time-invariant differences between municipalities. To address concerns about compositional changes we have estimated the baseline model with and without the set of covariates capturing important child and parent characteristics. We stress that our estimates should be interpreted as ITT-effects, and should be scaled by about 1.9 to arrive at the ATT.

Mean effect. Table 2 reports our difference-in-differences estimates based on equation (1) from the sample of children born 1990–1991. In Panel A, we report the estimated effects on exam performance at the end of compulsory school, with and without the set of covariates. The estimates indicate that the reform had little effect on children's school performance, with a precisely estimated point estimate of about 1 % of a standard deviation. Excluding covariates in the second row of Panel A hardly moves the estimate. This indicates that there are no important compositional changes between the two cohorts, as expected from historical reports and descriptive statistics. Given the precision of the estimate and scaling for take-up, we can rule out effects above 3.3 % and below -7.1 % of a standard deviation at a confidence level of 5 %.

While studies of how early interventions affect child cognitive outcomes often find positive effects in the short run, these effects are often found to dissipate over time (see e.g. Knudsen, Heckman, Cameron, and Shonkoff, 2006). At the same time, persistent effects are often found on outcomes that may also reflect non-cognitive traits. In Panels B and C of Table 3, we consider effects on high school dropout rates and enrollment in the academic track in upper secondary school, where earlier studies have often found an improvement from early intervention programs. However, again we find little evidence of any substantial effect, whether or not we include covariates. Indeed, if anything, we find a small negative impact on children's schooling of mandating kindergarten, with a slight rise in high school dropout rates of 1.3 percentage points (from a pre-reform mean of about 33 % in the treatment group).

Heterogeneous effects. Though we find little support for an effect of mandating kindergarten on mean school performance, high school dropout or choice of academic track, we have already emphasized the general expectation of heterogeneous effects of early childhood interventions. A worry may therefore be that estimated mean effects mask large but offsetting effects among different groups of children. One concern

Table 2: Mean effects on school performance, high school dropout rates and academic track in upper secondary school.

	Coeff	*SE*	*Mean [SD]*
A. School performance			
Baseline	-0.01	(0.014)	0 [1]
No covariates	-0.013	(0.016)	
B. High school dropout			
Baseline	0.013	(0.008)	0.33 [0.47]
No covariates	0.014	(0.008)	
C. Academic track			
Baseline	-0.008	(0.007)	0.40 [0.49]
No covariates	-0.009	(0.008)	

Note: $N = 111,397$ ($N = 107,707$ for academic track). Estimations are based on OLS on equation (1). The controls are listed in Table 1 and the dependent variables are defined in Sections 3 and 5. In Panel A, coefficients are standardized to the standard deviation of the dependent variable. Mean refers to pre-reform mean in the treatment group. Standard errors (SE) are robust for heteroskedasticity and all models include municipality fixed effects.

might be that effects would be offsetting over the distribution of the outcome, e.g. beneficial in the lower parts of the grading distribution but negative in the upper parts. Another concern might be that effects differ depending on characteristics of the child, the family or the local school, either due to heterogeneous responses, or due to different exposure to the treatment (i.e. different take-up rates).

To address the first concern, we have estimated the impact of the reform on school performance at every point in the grading distribution (Table A2). Specifically, we estimate equation (1) over the sample of children born 1990–1991, where the outcome variable is a dummy equal to one if the child's school performance at end of compulsory school is above the given percentile, and zero otherwise. Estimates should then be interpreted as the percentage point change following the reform in the probability of performing above a given percentile for a child in the treatment group compared to a child in the comparison group.[22] Estimates for selected percentiles covering the bottom, the middle and the top of the distribution are reported in Table A6 in the appendix.[23] Results show that there is essentially no heterogeneity across the grading distribution.

To address the second concern, we have estimated separate reform effects for all outcomes in subsamples defined from a number of background characteristics. The estimates are reported in Table 3. To facilitate comparison of estimates across subsamples, we also report the mean outcome among treated children from the pre-

[22]This procedure is essentially the first step in the RIF-procedure proposed by Firpo, Fortin, and Lemieux (2009) and applied to the DID-framework by Havnes and Mogstad (2012).

[23]We have estimated effects at all percentiles, which yields the same picture.

reform cohort, the share of treated children and the take-up rate in the subsample.[24] Note that there is no systematic relationship between the take-up and the estimated coefficient (see also Figure A2 in the appendix).

The results give little indication of important heterogeneity in the impact of mandatory kindergarten, which is estimated to be very small. However, lower sample size implies less precision, and some patterns in the point estimates may warrant comment. First, girls seem to benefit more than boys, in line with what is often found in the literature on cognitive impact of early childhood interventions (Anderson, 2008). Second, though very imprecise, we also note a pattern that children that initially perform well, as measured by mean exam grade pre-reform, may tend to receive the most harm from mandatory kindergarten. In particular, children of higher educated families on average do experience a modest negative effect of the reform. Though estimates are too imprecise to provide much confidence, this could be interpreted as an indication that parents with high levels of human capital provide a good alternative to kindergarten, in line with Cunha, Heckman, and Schennach (2010) and estimates in Havnes and Mogstad (2012).

6 Specification checks

To improve our confidence in the estimates, we now challenge our empirical approach in different ways. First, we confront the key identifying assumption of our empirical strategy, namely the common trend assumption. Second, we consider whether there may be a separate effect of the reform on our comparison group that may attenuate effects on our treatment group. Third, we investigate whether there might be a delayed effect of the reform on later cohorts, before we look at how the bedding-down of the new curriculum could threaten our estimates. Finally, we consider some alternative and less aggregated school outcomes to understand whether there may be effects on some particular sets of skills that are washed out in our aggregated measure, and investigate whether the reform may have had an effect on the labor supply of mothers. For brevity, we focus on school performance. Results are similar for high school dropout and enrollment in the academic track (cf. Tables A7 and A8 in the appendix).

Common trend assumption. The primary threat in DD estimation is that the change in the observed outcome in the comparison group in the absence of the reform differs

[24] As discussed in Section 4, take-up is defined as the probability that a child who does not enroll in child care at age five, and is therefore in our treatment group, does not enroll in kindergarten at age six, and should therefore be affected by the reform.

Table 3: Heterogeneous responses

	School performance			High school dropout			Academic track			Treated	Take-up
	Coeff	*SE*	*Mean*	*Coeff*	*SE*	*Mean*	*Coeff*	*SE*	*Mean*		
A. Child and family characteristics											
Girls	0.012	(0.020)	3.73	0.008	(0.009)	0.22	0.000	(0.011)	0.58	0.16	0.52
Boys	-0.032	(0.021)	3.44	0.017	(0.010)	0.26	-0.015	(0.011)	0.44	0.16	0.52
Single parent	-0.035	(0.029)	3.31	0.017	(0.015)	0.39	0.002	(0.015)	0.40	0.24	0.56
Low income	-0.008	(0.043)	3.30	-0.006	(0.022)	0.45	0.013	(0.022)	0.34	0.45	0.65
Young mother	-0.035	(0.034)	3.31	0.019	(0.018)	0.37	0.002	(0.018)	0.35	0.24	0.53
On welfare	-0.006	(0.038)	3.22	-0.001	(0.020)	0.43	0.014	(0.019)	0.34	0.35	0.63
B. Mother characteristics											
No high school	0.003	(0.019)	3.35	0.020	(0.010)	0.41	-0.007	(0.010)	0.33	0.22	0.57
High school	-0.021	(0.023)	4.00	-0.003	(0.010)	0.21	-0.012	(0.012)	0.59	0.11	0.40
College	-0.047	(0.036)	4.28	0.014	(0.014)	0.14	0.003	(0.018)	0.72	0.08	0.32
C. Father characteristics											
No high school	0.016	(0.021)	3.31	0.016	(0.011)	0.43	-0.002	(0.011)	0.29	0.20	0.55
High school	-0.025	(0.020)	3.84	0.005	(0.008)	0.25	-0.013	(0.010)	0.51	0.14	0.48
College	-0.021	(0.032)	4.18	0.007	(0.012)	0.17	-0.029	(0.016)	0.70	0.12	0.47

Note: $N = 111,397$ ($N = 107,707$ for academic track). Estimations are based on OLS on equation (1). The subsamples are defined in Section 3. The controls are listed in Table 1 and the dependent variables are defined in Section 3. Standard errors (SE) are robust for heteroskedasticity and all models include municipality fixed effects.

from the change in the potential outcome of the treatment group in the absence of the reform. One example could be anticipation effects, e.g. knowing that kindergarten would be free from age six, children born in 1991 might be more likely to be enrolled at age five. To investigate the common trend assumption, we start by considering a placebo reform, pretending that the reform was implemented in the pre-reform period. The first row of Panel A in Table 4 reports the estimate from equation (1) estimated over the sample of children born 1989–1990, where $Post_t$ is redefined to be equal to one for children born in 1990 and zero otherwise. A significant estimate in this specification would put in doubt our identifying assumption. However, the estimate is almost precisely zero and nowhere near statistical significance.

Allowing treatment and comparison groups to follow separate trends is another way to challenge the common trend assumption. By extrapolating pre-reform trends into the post-reform period, we essentially restrict our estimates to reflect how outcomes deviate from the pre-reform trajectory. As emphasized by Besley and Case (2000), this is a simple yet potentially powerful test, which can often kill otherwise large and significant DD estimates.

To allow estimation of a trend, we extend the estimation sample to the start of our data series in 1988, and include the 1992-cohort, which is the last cohort that we can confidently use due to the cash-for-care reform in 1998. We then set $Post_t = 1$ for $t = 1991$ and $t = 1992$, and zero otherwise. For a correct comparison, results on this sample using our main regression, equation (1), are reported in the first row of Panel B. Estimates conform to those in the baseline. In row 2 of Panel B, we include a linear treatment-specific trend, while row 3 includes a second-order polynomial treatment-specific trend. Both specifications confirm the baseline estimates of essentially no effect of introducing mandatory kindergarten.

As an alternative, we can instead follow Duflo (2001) in allowing children to follow different trends depending on underlying characteristics. Specifically, we first estimate equation (1) including a linear trend interacted with baseline covariates.[25] We then relax the assumption of a linear trend, interacting instead the baseline covariates with the cohort fixed effects. Results are reported in Panel C, again confirming our baseline estimate of hardly any impact of the mandatory kindergarten reform on children's school performance at end of compulsory schooling.

Effect on comparison group. The transfer of kindergarten teachers from child care centers into schools, meant that six year olds in 1997 were likely to experience a similar pedagogical environment to the one they would have experienced in the

[25]The baseline covariates are measured the year the child turns five years old, and include an overall measure of school size, the education level of the mother and father, the average income in the municipality of residence, and a dummy indicating whether the child lives in an urban area.

Table 4: Robustness – School performance

	Sample	Post	Coeff	SE	N
A. Key specification check					
Placebo	1989–1990	1990	0.006	(0.016)	110,171
B. Treatment-specific trends					
Extending pre-reform	1988–1992	1991–92	-0.007	(0.010)	267,745
Linear trend	1988–1992	1991–92	-0.017	(0.017)	267,745
Quadratic trend	1988–1992	1991–92	-0.008	(0.022)	267,745
C. Flexible trends					
Trend × covar	1988–1992	1991–92	0.016	(0.009)	267,745
Year FE × covar	1988–1992	1991–92	0.009	(0.009)	267,745
D. Other					
1st diff.: Treatment	1990–1991	1991	-0.006	(0.013)	18,108
1st diff.: Comparison	1990–1991	1991	0.004	(0.008)	93,288
Delayed effect	1988–1992	1991	-0.004	(0.011)	267,745
		1992	-0.011	(0.014)	

Note: Column 2 gives the estimation sample. In all estimations, $Post_t = 1$ for t is given in Column 3. In Panel A estimation is based on OLS on equation (1). In Panel B, estimations are based on equation (1), including a linear (row 2) and a quadratic (row 3) treatment-specific trend. In Panel C, estimations are based on equation (1), including a linear trend (row 1) or cohort dummies (row 2) interacted with a set of baseline covariates (school size; mother's and father's education level; municipal income; urban area). In rows 1 and 2 of Panel D, estimations are based on equation (2), while row 3 is based on equation (3). The controls are listed in Table 1 and the dependent variables are defined in Section 3. Standard errors (SE) are robust for heteroskedasticity and all models include municipality fixed effects.

absence of the reform. Even so, the potential for changes in the content of the program coinciding with the reform raises the concern that the reform could have had an effect also on children in the comparison group. If the kindergarten program offered prior to the reform was of lower quality than the program offered after the reform, then the comparison group would experience a positive impact of the reform. In this case, the new mandatory kindergarten program may in fact have a positive effect for children in the treatment group that is simply netted out in our DD-setup against a positive effect (of similar size) in the comparison group. Similarly, a negative impact on the comparison group could mask a negative impact in the treatment group. To investigate this, we consider the two groups of children separately, to reveal whether there are in fact substantial changes in the grades of children around the implementation of the reform. Looking back at Panel (b) of Figure 2, we see no indication of such changes in neither the treatment nor the comparison group. More formally, and including covariates, we estimate first

difference regressions separately for the two groups based on

$$Y_{it} = \alpha + \lambda Post_t + X_{it}^{'}\beta + \epsilon_{it} \tag{2}$$

Results reported in rows 1–2 in Panel D of Table 4 give no reason to believe that mandating kindergarten had much impact on neither the comparison group, nor the treatment group. *Bedding down effect.* Children in our main sample, born 1990–1991 and entering grades 1 and 2 in 1997, started school in the same year, and were subject to the same, new curriculum throughout primary school. Children born in 1990 were the first to experience a new curriculum for grades 2–4, while children born in 1991 were the first to experience a new curriculum in grade 1, but the second to experience a new curriculum for grades 2–4. If the bedding down of the curriculum was a disadvantage for the children born in 1990, then our estimates could be biased upwards, and might mask an otherwise negative effect. To investigate this, we first note that if bedding down effects are important, then we would expect the performance of the 1990-cohort to dip compared to the 1989 cohort. Figure 2 shows no evidence of such a dip. Of course, our estimates would not be affected by such overall effects in any case, since our DD estimation strategy removes cohort effects. We may, however, worry that children who did not attend kindergarten before starting school (our treatment group) are more sensitive to the bedding down of the curriculum. To investigate this, we take advantage of the fact that the 1989 cohort were never exposed to the new curriculum as a first cohort. If bedding down effects are important and affect our treated group disproportionately, then we would expect a negative DD-effect when we compare the 1989 and 1990-cohort. Again, we see no evidence of this in Figure 2. Also, this is precisely what is estimated in the placebo test in Table 4, where the estimate is essentially zero.

We have also performed an additional placebo estimation, analogous to the previous, where we compare the 1988 cohort to the 1989 cohort. These cohorts were both transferred to the new curriculum in 1999, and the difference in exposure is therefore modest. Again, we find no evidence of diverging trends (b=0.008, SE=0.015).

Delayed effect. A further worry may be that a positive effect of the mandatory kindergarten reform was offset, completely or in part, by adjustment problems in the year of implementation. Unfortunately, the cash-for-care reform implemented in late 1998 (see Section 2), creates problems for identifying effects on cohorts born in 1993 and onwards. We can, however, plausibly estimate effects on children born in 1992. To create balance between the comparison group and the treatment group, and to provide better identification of control variables, we also use the extended sample

of children born 1988–1992. To allow for different treatment effects on children born in the two years, we expand on equation (1) to include a separate interaction term for the 1992-cohort, i.e.

$$\begin{aligned} Y_{it} &= \alpha_t + \gamma_1 Treated_i + \lambda_{91} Treated_i \times 1\,(t = 1991) \\ &\quad + \lambda_{92} Treated_i \times 1\,(t = 1992) + X_{it}^{'}\beta + \epsilon_{it} \end{aligned} \tag{3}$$

where $1\,(t = s)$ is an indicator equal to one if $t = s$ and zero otherwise. Results are reported in the final row of Panel D, again revealing no evidence that the reform had an important impact on children's school performance at end of compulsory schooling. If anything, point estimates indicate that the 1992 cohort was doing worse than the 1991 cohort, suggesting that there was no delayed benefit of the program.

Alternative outcomes. We may also worry that the dependent variable is not picking up the relevant margin of the effect. For instance, if kindergarten affects mostly oral skills or mostly skills that are relevant in one or a few particular subjects, then our estimate may be small simply because it is diluted by including subjects in our outcome that test skills that are not affected. To investigate this, we have also considered alternative outcomes that should reflect different sets of skills. For brevity, estimates are reported in Table A9 in the appendix.

We start by separating the written and oral exams that make up our main dependent variable. Since there are usually two written exams for each oral exam, any effect on the oral exam may be diluted by a zero or counteracting effect on the written exam. In Panel A of Table A9, we report results from our baseline regression where the dependent variable is replaced by first the average of written exam grades and then by the grade on the oral exam. Estimated effects are virtually identical. Next, we consider teacher-assigned grades at end of compulsory school, available for 13 subjects (see Section 3). For comparison with our main estimates, we first run our baseline regression on the overall grade point average (GPA; the mean grade across all subjects). Not surprisingly, the estimated effect on the overall GPA is virtually identical to the effect on the average exam grade used in our main analysis (cf. Panel B of Table A9). We then separate out the subjects that are tested on the written and oral exams used in our main analysis, and those that are not tested on these exams.[26] Again, estimates are virtually identical to the baseline.

Finally, we group subjects according to the types of skills expected to deter-

[26]Subjects tested are written and oral Norwegian, written and oral English, mathematics, natural science, social science and religion. Subjects not tested are home economics, physical education, music, and arts and crafts.

mine the performance. Specifically, we group subjects into the following categories: "Sciences" (mathematics, natural science, and social science), "Languages" (written and oral Norwegian, and written and oral English), and "Culture" (religion, music, home economics and arts and crafts). Estimates are reported in rows 4–6 in Panel B of Table A9, and are again virtually identical. We conclude, therefore, that there is no evidence of substantial effects that were not picked up in our main analysis.

Income effects. There are two alternative channels for income effects: one from a change in the use or price of child care/after school programs, and another from a change in parental labor supply following the reform. For children that would attend child care before the reform and the after school program after the reform, the price difference is relatively small (cf. p. 30). For children that would not attend child care before the reform and not attend the after school program after the reform, the price is always zero. Finally, some children who would not attend child care before the reform may choose to opt into the after-school program after the reform. These children would experience an increase in the price, and hence a negative income effect that could adversely affect their children. We believe that this latter effect is small for two reasons. First, parents who opt out of voluntary kindergarten should be likely also to opt out of after-school care. Second, parents who choose to opt into the after-school program would likely be close to indifferent between attending kindergarten in the first place. This may suggest that the price is not very high compared to their income.

To investigate effects on parental labor supply, we have estimated the impact of the mandatory kindergarten reform on maternal labor supply. This is important in itself, but could also indirectly affect child performance by increasing family income (Dahl and Lochner, 2012; Løken, Mogstad, and Wiswall, 2010).[27] Following standard practice, we restrict our analysis to mothers with their youngest child born in 1990 or 1991, where we would expect the strongest labor supply responses. We apply an analogous empirical strategy to the one in our main analysis, estimating the DD-model in equation (1) where t refers to the cohort of the youngest child, $Treat$ is equal to one for mothers of children who enroll in child care at age five, while $Post$ is equal to one if the youngest child is born in 1991 and zero if the child

[27]In a survey of the early literature, Blau and Currie (2006) report elasticities of maternal employment with respect to the price of child care ranging from 0 to -1. More recently, using more plausible identification, Baker, Gruber, and Milligan (2008) find a positive effect on maternal labor supply following the introduction of heavily subsidized universally available child care in Quebec. Meanwhile, Lundin, Mork, and Ockert (2008) find no such effect when studying a childcare reform which capped childcare prices in Sweden. See also Schlosser (2005); Cascio (2009); Havnes and Mogstad (2011); Lefebvre and Merrigan (2008a) and Berlinski and Galiani (2007). For a review of the literature, see Blau and Currie (2006).

is born in 1990.[28] Results reveal that the reform had no impact on the labor market attachment of mothers in our sample (cf. Panel C of Table A5).

7 Concluding remarks

Evidence on the impact of child care interventions has been dominated by estimates from targeted programs. These may be hard to apply to the general population. Recent research provides some insight into the effects of large-scale programs. While the literature is expanding rapidly, there is still a lot we do not know about the impact of the universal programs advocated in many western countries (Baker, 2011). This is particularly worrisome given the heterogeneity created by wide differences in individual alternatives to subsidized care. The high returns found for children from disadvantaged families, coupled with much lower participation rates in existing programs compared to children from more advantaged backgrounds, suggests a potentially strong social gradient in expanding or mandating early childhood interventions (Barnett and Belfield, 2006). Indeed, in an effort to counter differences at school entry depending on social background, many countries are currently moving towards subsidized kindergarten or child care available for the general population.

In the current paper, we provide first evidence on the effect of mandating kindergarten at age 5–6 on children's schooling outcomes. Specifically, we consider the impact on school performance at the end of compulsory schooling at age 15–16, on high school dropout and on the likelihood of enrolling in an academic track. Our identifying variation comes from a 1997-reform in Norway that lowered school starting age from seven to six. The goal of the reform was to counter differences in learning outcomes between children from different socioeconomic backgrounds. The contents and structure of the program bear resemblance to the U.S. Head Start program and to the early U.S. kindergarten program, in focussing mostly on social development and less on the aquisition of academic skills.

Our results reveal that the reform did little to counter differences in schooling outcomes between children from different socioeconomic backgrounds. In our baseline estimation, the precisely estimated effect on the child's exam performance is below 2 % of a standard deviation. Estimates are similarly small when we consider effects across the grading distribution and in different subsamples defined from char-

[28] As our outcome of interest, we consider labor supply of mothers in the year when the child turns seven years old. That is, ideally we consider the labor supply of mothers whose youngest child was enrolled or was not enrolled in kindergarten in the months January through late August. We have also estimated effects when the child is six (when the child may be enrolled September through December), finding no impact of the reform.

acteristics of the child or parents. A number of specification checks lend support to our empirical strategy. A lack of effects on medium-term school performance could still mask a long-term effect on substantive outcomes. Evidence from the Perry Preschool program suggests that although an initial increase in IQ among treated children faded out and effects on school performance for boys were at best modest, the program still generated positive effects on longer-term outcomes such as crime and labor market behavior (Heckman, Moon, Pinto, Savelyev, and Yavitz, 2010; Heckman, Pinto, and Savelyev, 2013). However, we also find negligible impacts on high school dropout (age 18) and academic tracking (age 16) which may arguably be of direct substantive interest.

While the program we study is universal, the reform may be viewed as targeted since the affected children come disproportionately from disadvantaged families. One may therefore interpret our results as shedding light on how a universal low intensity program can improve outcomes among the disadvantaged. Previous evidence suggests that low intensity child care has the potential to improve child outcomes in this group. Our evidence suggests that this may no longer be the case when the program is truly universal. This may not necessarily come as a surprise, since the larger operations and broader scope involved in a universal compared to a targeted program may come with particular challenges, for instance by making it harder to see children's needs and to tailor activities to these needs. One interpretation may then be that the unstructured child-centered approach to instruction (Stipek, Feiler, Daniels, and Milburn, 1995), which has been a hallmark of low intensity child care programs, may be less suitable in a universal program. Our study may then lend support to policies aimed at improving the intensity of targeted programs over policies aimed at expanding the reach of these programs. Another interpretation may be that parents are able to sort children relatively efficiently into child care programs based on their children's individual needs, suggesting that mandating participation may not be effective in reducing socioeconomic differences between children. This might be particularly true when child care programs prior to mandating are widely available and affordable, as in our case. We emphasize that our analysis is based on children from non-immigrant families, and does not, therefore, speak to the debate on early interventions to provide language training among non-native speakers.

We believe that our evidence along with e.g. Baker, Gruber, and Milligan (2008), may call for caution in the current push towards using universal child care as a tool to promote the development of children from disadvantaged families. For instance, it could look as if the European Union Commission is reading too much into descriptive, rather than causal, evidence when they proclaim that "[t]here is clear evidence

that universal access to quality ECEC is more beneficial than interventions targeted exclusively at vulnerable groups" (European Union, 2011, p. 5). While we agree that early childhood investments can be an important tool in facilitating equal opportunities, our evidence emphasizes that this is hardly automatic, and suggests that the structuring of the program and its content may be key to generating the intended benefits.

Finally, the conclusion that mandating kindergarten had little impact on children's school performance may cut both ways. While the large benefits expected by proponents can be firmly rejected, our results also lend little support to claims of strong negative effects from opponents. This is true even though the reform implemented a fully mandated program affecting families that did not voluntarily enroll their children, and who would otherwise care for their children themselves. It should be noted, however, that these estimates are driven mostly by children from relatively lower socioeconomic backgrounds, and may not be representative for children from higher socioeconomic backgrounds. Also, it is clear that the evidence on how universal or large-scale child care affects child development is mixed and still quite scarce, and that one reason could be that the alternative mode of care differs across countries. It is therefore of great importance to accumulate more evidence on how child care programs can affect child development.

References

Almond, D., and J. Currie (2011): *Human Capital Development before Age Five*vol. 4 of *Handbook of Labor Economics*, chap. 15, pp. 1315–1486. Elsevier.

Anderson, M. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Intervention Training Projects.," *Journal of the American Statistical Association*, 103(484), 1481–1495.

Atkinson, A. B., L. Rainwater, and T. M. Smeeding (1995): *Income distribution in OECD countries : evidence from the Luxembourg Income Study*. OECD Publications and Information Center, Paris.

Baker, M. (2011): "Innis Lecture: Universal early childhood interventions: what is the evidence base?," *Canadian Journal of Economics*, 44(4), 1069–1105.

Baker, M., J. Gruber, and K. Milligan (2008): "Universal Child Care, Maternal Labor Supply, and Family Well-Being," *The Journal of Political Economy*, 116(4), pp. 709–745.

Barnett, W. S. (1995): "Long-term effects of early childhood programs on cognitive and school outcomes," *Future of Children*, pp. 22–50.

Barnett, W. S., and C. R. Belfield (2006): "Early Childhood Development and Social Mobility," *The Future of Children*, 16(2), 73–98.

Berlinski, S., and S. Galiani (2007): "The effect of a large expansion of pre-primary school facilities on preschool attendance and maternal employment," *Labour Economics*, 14(3), 665–680.

Berlinski, S., S. Galiani, and M. Manacorda (2008): "Giving children a better start: Preschool attendance and school-age profiles," *Journal of Public Economics*, 92(5-6), 1416–1440.

Besley, T., and A. Case (2000): "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," *Economic Journal*, 110(467), F672–94.

Bettinger, E., T. Hægeland, and M. Rege (2014): "Home with Mom: The Effects of Stay-at-Home Parents on Children's Long-Run Educational Outcomes," *Journal of Labor Economics*, 32(3), 443–467.

Black, S. E., P. J. Devereux, K. V. Løken, and K. G. Salvanes (2014): "Care or cash? The effect of child care subsidies on student performance," *Review of Economics and Statistics*, 96(5), 824–837.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2011): "Too Young to Leave the Nest? The Effects of School Starting Age," *The Review of Economics and Statistics*, 93(2), 455–467.

Blau, D., and J. Currie (2006): *Pre-School, Day Care, and After-School Care: Who's Minding the Kids?*vol. 2 of *Handbook of the Economics of Education*, chap. 20, pp. 1163–1278. Elsevier.

Cascio, E. U. (2009): "Do Investments in Universal Early Education Pay Off? Long-term Effects of Introducing Kindergartens into Public Schools," Working Paper 14951, National Bureau of Economic Research.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78(3), 883–931.

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

Currie, J., and D. Thomas (1995): "Does Head Start make a Difference?," *American Economic Review*, 85(3), 341–364.

Dahl, G. B., and L. Lochner (2012): "The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit," *American Economic Review*, 102(5), 1927–56.

DeCicca, P., and J. Smith (2013): "The Long-Run Impacts of Early Childhood Education: Evidence from a Failed Policy Experiment," *Economics of Education Review*, 36, 41–59.

Deming, D. (2009): "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start," *American Economic Journal: Applied Economics*, 1(3), 111–134.

Drange, N., and M. Rege (2013): "Trapped at home: The effect of mothers' temporary labor market exits on their subsequent work career," *Labour Economics*, 24, 125–136.

Duflo, E. (2001): "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4), 795–813.

Dustmann, C., A. Raute, and U. Schonberg (2013): "Does Universal Child Care Matter? Evidence from a Large Expansion in Pre-School Education," Mimeo.

European Union (2011): *Early Childhood Education and Care: Providing all our children with the best start for the world of tomorrow.*

Felfe, C., and R. Lalive (2013): "Early Child Care and Child Development: For Whom it Works and Why," CEPR Discussion Papers 9274, C.E.P.R. Discussion Papers.

Felfe, C., N. Nollenberger, and N. Rodríguez-Planas (2012): "Can't Buy Mommy's Love? Universal Childcare and Children's Long-Term Cognitive Development," *Journal of population economics*, 28(2), 393–422.

Firpo, S., N. M. Fortin, and T. Lemieux (2009): "Unconditional Quantile Regressions," *Econometrica*, 77(3), 953–973.

Fitzpatrick, M. D. (2008): "Starting School at Four: The Effect of Universal Pre-Kindergarten on Children's Academic Achievement," *The B.E. Journal of Economic Analysis & Policy*, 8(1).

Garces, E., D. Thomas, and J. Currie (2000): "Longer Term Effects of Head Start," Working Paper 8054, National Bureau of Economic Research.

Gupta, N. D., and M. Simonsen (2010): "Non-cognitive child outcomes and universal high quality child care," *Journal of Public Economics*, 94(1-2), 30 – 43.

Havnes, T., and M. Mogstad (2011): "No Child Left Behind. Subsidized Child Care and Children's Long-Run Outcomes," *American Economic Journal: Economic Policy*.

——— (2012): "Is Universal Child Care Leveling the Playing Field?," CESifo Working Paper Series 4014, CESifo Group Munich.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *The American Economic Review*, 103(6), 2052–2086.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010): "The rate of return to the HighScope Perry Preschool Program," *Journal of public Economics*, 94(1), 114–128.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–75.

Karoly, L. A., M. R. Kilburn, and J. S. Cannon (2005): *Early Childhood Interventions: Proven Results, Future Promise*. RAND Corporation, Santa Monica, CA.

Knudsen, E. I., J. J. Heckman, J. L. Cameron, and J. P. Shonkoff (2006): "Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce," *Proceedings of the National Academy of Sciences of the United States of America*, 103(27), pp. 10155–10162.

Lefebvre, P., and P. Merrigan (2008a): "Child-Care Policy and the Labor Supply of Mothers with Young Children: A Natural Experiment from Canada," *Journal of Labor Economics*, 26(3), 519–548.

——— (2008b): "Family Background, Family Income, Cognitive Tests Scores, Behavioural Scales and their Relationship with Post-secondary Education Participation: Evidence from the NLSCY," Cahiers de recherche 0830, CIRPEE.

Leuven, E., M. Lindahl, H. Oosterbeek, and D. Webbink (2010): "Expanding schooling opportunities for 4-year-olds," *Economics of Education Review*, 29(3), 319–328.

Løken, K. V., M. Mogstad, and M. Wiswall (2010): "What Linear Estimators Miss: Re-Examining the Effects of Family Income on Child Outcomes," IZA Discussion Papers 4971, Institute for the Study of Labor (IZA).

Ludwig, J., and D. L. Miller (2007): "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *The Quarterly Journal of Economics*, 122(1), 159–208.

Lundin, D., E. Mork, and B. Ockert (2008): "How far can reduced childcare prices push female labour supply?," *Labour Economics*, 15(4), 647–659.

McKey, R. H., L. Condelli, H. Ganson, B. J. Barrett, C. McConkey, and M. Plantz (1985): "The Impact of Head Start on Children, Families and Communities. Final Report of the Head Start Evaluation, Synthesis and Utilization Project.," Discussion paper, U.S. Department of Health and Human Services.

Norwegian Ministry of Child and Family Affairs (1995): *Lov om barnehager. Rundskriv Q-0902B. (The Child Care Act).*

Norwegian Ministry of Education (1990-91): *Ot.prp. nr. 57: Om lov om endring av lov 6.juni 1975 nr. 30 om barnehage m.m.,.*

——— (1992-93): *St.meld. nr. 40: ...vi smaa, en Alen lange; Om 6-åringer i skolen - konsekvenser for skoleløpet og retningslinjer for dets innhold.*

——— (1993-94): *Innst. O. nr. 36: Innstilling fra kirke-, utdannings- og forskningskomiteen om lov om endringer i lov av 13. juni 1969 nr. 24 om grunnskolen.*

——— (1996a): *L97: Læreplanverket for den 10-årige grunnskolen.*

——— (1996b): *Reform 97: Dette er grunnskolereformen.*

——— (2010): *NOU 2010: 8: Med forskertrang og lekelyst. Systematisk pedagogisk tilbud til alle førskolebarn (Systematic educational offering for preschool children)*, no. 8.

Reardon, S. F. (2011): "The widening academic achievement gap between the rich and the poor: New evidence and possible explanations," *Whither opportunity*, pp. 91–116.

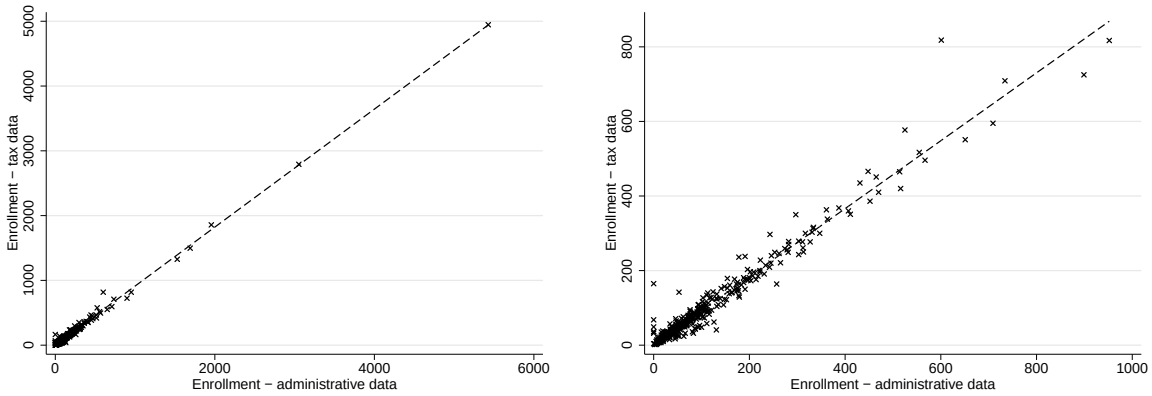
Ruhm, C., and J. Waldfogel (2012): "Long-term effects of early childhood care and education," *Nordic Economic Policy Review*, 1(1), 23–51.

Schlosser, A. (2005): "Public Preschool and the Labor Supply of Arab Mothers: Evidence from a Natural Experiment," Discussion paper, Mimeo, The Hebrew University of Jerusalem.

Schøne, P. (2004): "Labour supply effects of a cash-for-care subsidy," *Journal of Population Economics*, 17(4), 703–727.

Stipek, D., R. Feiler, D. Daniels, and S. Milburn (1995): "Effects of Different Instructional Approaches on Young Children's Achievement and Motivation," *Child Development*, 66, 209–223.

A Appendix



(a) Full sample (b) Truncated sample

Figure A1: Enrollment – tax data and administrative data

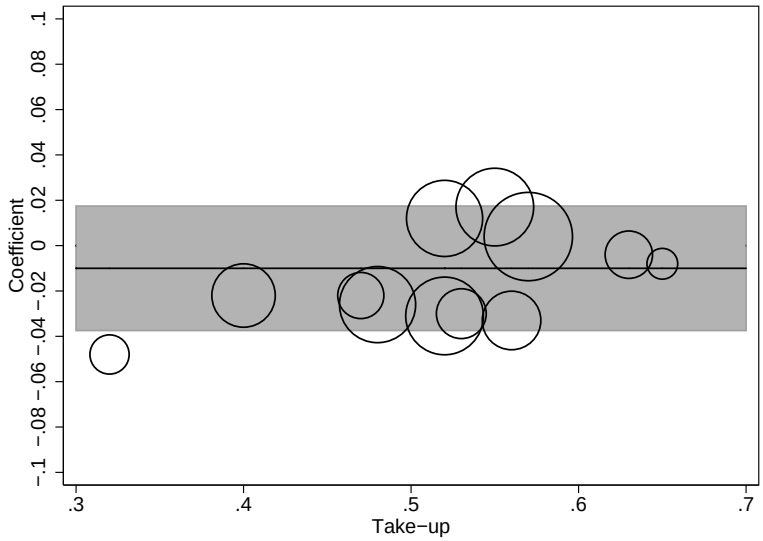


Figure A2: Estimated reform effect and take-up across subsamples

Note: The horizontal line and the shaded area correspond to the baseline estimate and its 95% confidence interval. Circles mark the estimates and take-up rates for subgroups reported in Table 3. The size of the circle indicates the size of the subgroup among treated children born in 1990. Take-up is defined as the probability that a child born in 1990 and not enrolled in child care at age five was also not enrolled in kindergarten at age six, see Section 4. Coefficients and take-up rates for subgroups are reported in Table 3.

Table A5: School performance and enrollment in kindergarten by family income decile at age five, children born in 1990.

Decile	Family income	School performance	Enrollment in kindergarten
1	15,912 (12,640)	3.42 (1.00)	0.54 (0.50)
2	45,124 (4,462)	3.53 (0.98)	0.65 (0.48)
3	57,408 (2,962)	3.60 (0.97)	0.72 (0.45)
4	66,968 (2,576)	3.68 (0.97)	0.82 (0.38)
5	75,547 (2,380)	3.73 (0.95)	0.88 (0.33)
6	83,524 (2,271)	3.79 (0.93)	0.91 (0.29)
7	91,695 (2,462)	3.86 (0.95)	0.92 (0.27)
8	101,363 (3,249)	4.02 (0.93)	0.93 (0.26)
9	116,100 (5,757)	4.11 (0.91)	0.93 (0.26)
10	169,179 (90,272)	4.30 (0.89)	0.92 (0.27)

Note: This table corresponds to Figure 1. School performance is measured as the average exam performance at the end of compulsory schooling (age 15–16). Enrollment in kindergarten is measured at age five. Standard deviations are in parentheses.

Table A6: Distributional effects on school performance

	Coeff	SE	Perc. value
5th percentile	0.002	(0.004)	2.0
10th percentile	-0.001	(0.006)	2.5
25th percentile	-0.011	(0.007)	3.0
50th percentile	-0.001	(0.006)	4.0
75th percentile	-0.005	(0.005)	4.5
90th percentile	-0.005	(0.004)	5.0
95th percentile	-0.002	(0.002)	5.5

Note: $N = 111,397$. Estimations are based on OLS on equation (1). The controls are listed in Table 1 and the dependent variable is defined in Section 3 and 5. Percentile values refer to pre-reform percentiles in the treatment group. Standard errors (SE) are robust for heteroskedasticity and all models include municipality fixed effects.

Table A7: Robustness – High school drop out

	Sample	Post	Coeff	SE	N
A. Key specification check					
Placebo	1989–1990	1990	-0.001	(0.007)	110,171
B. Treatment-specific trends					
Extended sample	1988–1991	1991	0.013	(0.005)	218,485
Linear trend	1988–1991	1991	0.011	(0.009)	218,485
Quadratic trend	1988–1991	1991	0.019	(0.023)	218,485
C. Flexible trends					
Trend × covar	1988–1991	1991	0.005	(0.006)	213,472
Year FE × covar	1988–1991	1991	0.008	(0.006)	213,472
D. Other					
1st diff.: Treatment	1990–1991	1991	0.017	(0.007)	18,108
1st diff.: Comparison	1990–1991	1991	0.004	(0.003)	93,288

Notes: Column 2 gives the estimation sample. In all estimations, $Post_t = 1$ for t is given in Column 3. In Panel A, estimation is based on OLS on equation (1). In Panel B, estimations are based on equation (1), including a linear (row 2) and a quadratic (row 3) treatment-specific trend. In Panel C, estimations are based on equation (1), including a linear trend (row 1) or cohort dummies (row 2) interacted with a set of baseline covariates (school size; mother's and father's education level; municipal income; urban area). In rows 1 and 2 of Panel D, estimations are based on equation (2), while row 3 is based on equation (3). The controls are listed in Table 1 and the dependent variables are defined in Section 3. Standard errors (SE) are robust for heteroskedasticity and all models include municipality fixed effects.

Table A8: Robustness – Academic track

	Sample	Post	Coeff	SE	N
A. Key specification check					
Placebo	1989–1990	1990	-0.003	(0.008)	105,894
B. Treatment-specific trends					
Extending pre-reform	1988–1992	1991–92	-0.003	(0.005)	258,112
Linear trend	1988–1992	1991–92	-0.008	(0.009)	258,112
Quadratic trend	1988–1992	1991–92	-0.016	(0.011)	258,112
C. Flexible trends					
Trend × covar	1988–1992	1991–92	0.005	(0.005)	252,495
Year FE × covar	1988–1992	1991–92	0.002	(0.005)	252,495
D. Other					
1st diff.: Treatment	1990–1991	1991	-0.020	(0.007)	17,203
1st diff.: Comparison	1990–1991	1991	-0.011	(0.003)	90,504
Delayed effect	1988–1992	1991	-0.009	(0.006)	258,112
		1992	0.005	(0.007)	

Note: Column 2 gives the estimation sample. In all estimations, $Post_t = 1$ for t is given in Column 3. In Panel A, estimation is based on OLS on equation (1) for years 1989 and 1990, years 1988 and 1989 and years 1988 and 1990. In Panel B, estimations are based on equation (1), including a linear (row 2) and a quadratic (row 3) treatment-specific trend. In Panel C, estimations are based on equation (1), including a linear trend (row 1) or cohort dummies (row 2) interacted with a set of baseline covariates (school size; mother's and father's education level; municipal income; urban area). In rows 1 and 2 of Panel D, estimations are based on equation (2), while row 3 is based on equation (3). The controls are listed in Table 1 and the dependent variables are defined in Section 3. Standard errors (SE) are robust for heteroskedasticity and all models include municipality fixed effects.

Table A9: Alternative outcomes

	Coeff	SE	N	Mean
A. Separating written and oral exams				
Exam, written subjects	-0.012	(0.015)	108,473	3.21
Exam, oral subjects	-0.012	(0.015)	105,224	4.05
B. Teacher-assigned grades				
Grade point average	-0.013	(0.014)	111,185	3.79
Exam subjects	-0.014	(0.014)	111,021	3.64
Non-exam subjects	-0.007	(0.015)	111,038	4.12
Sciences	-0.011	(0.014)	111,225	3.57
Languages	-0.015	(0.015)	110,951	3.66
Culture	-0.008	(0.014)	111,157	4.11
C. Mothers labor supply, child age 7				
Earnings (2011-USD)	1,540	(914)	46,742	20,110
Employment	-0.002	(0.010)	46,742	0.23
Full time	-0.000	(0.008)	46,742	0.10

Note: In Panel A and B estimations are based on OLS on equation (1) including covariates. "Sciences" includes mathematics, natural science, and social science; "Languages" includes written and oral Norwegian, and written and oral English; "Culture" includes religion, music, home economics and arts and crafts. In Panel C estimations are based on OLS on equation (1), and the sample is restricted to mothers with youngest child of relevant age. We define a mother as being employed if she is registered with more than four working hours per week, and in full time employment if she is registered with more than 30 working hours per week while earning more than two times the basic amount in the Norwegian pension system (about USD 26,000). We restrict full time employment also on the level of earnings to correct for lags in the submission of employee information by firms, which causes some individuals with low or even zero earnings to be recorded as full time workers. The basic amount of the Norwegian Social Insurance Scheme is used to define labor market status, and determine eligibility for unemployment benefits as well as disability and old age pension. Covariates included are listed in Table 1 (in Panel C, we exclude measures of mothers employment and include municipality-specific unemployment rates). Standard errors (SE) are robust for heteroskedasticity and all models include municipality fixed effects.

Chapter 2:

Do smaller classes always improve students' long-run outcomes?

Torberg Falch, Astrid Marie Jorde Sandsør and Bjarne Strøm

Do smaller classes always improve students' long-run outcomes?*

Torberg Falch[†] Astrid Marie Jorde Sandsør[‡] Bjarne Strøm[§]

Abstract

We exploit the strict class size rule in Norway and matched individual and school register information for 1982−2011 to estimate long-run causal effects on income and educational attainment. Contrary to recent evidence from the United States and Sweden, we do not find any significant average effect on long-run outcomes of reduced class size. We use the large register data set and quasi-experimental strategy to estimate whether the class size effect depends on external conditions facing students and schools, such as teacher quality, extent of upper secondary school choice, school district size, local fiscal constraints, and labor market conditions. Overall, we find that the class size effect does not depend on these factors measured at the school district level. The absence of class size effects on long-run outcomes in Norway is consistent with earlier findings for short-run outcomes using comparable data and empirical strategies.

Keywords: class size, school district, quasi-experiment, educational attainment, income
JEL codes: I2, H7

*We greatly acknowledge comments from Kalle Moene and participants at the Workshop on Applied Economics of Education in Catanzaro, the CESifo Area Conference on Economics of Education, and seminars at University of Lancaster and the Institute of Social Research in Oslo.

[†]Department of Economics, Norwegian University of Science and Technology
[‡]Department of Economics, University of Oslo
[§]Department of Economics, Norwegian University of Science and Technology

1 Introduction

The impact of school resources on student performance has been disputed since the publication of the Coleman report (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York, 1966). Although availability of data and empirical strategies to uncover causal effects have increased substantially in recent years, the evidence on the effect of resources on education outcomes is still inconclusive.[1] The literature is not conclusive even for more narrow and popular policy tools as class size. Although the results from the well known randomized Student/Teacher Achievement Ratio experiment (Project STAR) in Tennessee suggest that smaller classes are beneficial in terms of test scores,[2] studies using quasi-experimental approaches to identify causal effects differ substantially in their conclusions.[3] One interpretation is that extra resources and reduced class size are effective tools in some contexts, while ineffective in other contexts.

Academic test scores only measure cognitive skills, while class size may also affect non-cognitive skills. In addition, evidence based on test scores may be biased in settings where teachers systematically manipulate test scores as recently demonstrated in Angrist, Battistin, and Vuri (2015).[4] Both arguments suggest that analyses of long-run outcomes in terms of educational attainment and income in adulthood as used in our empirical study would provide the most credible evidence of the effect of school resources. Such studies will embed all short-run effects, including effects on non-cognitive skills that are difficult to measure directly.

Three recently published papers analyze long-run effects of class size. Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) and Dynarski, Hyman, and Schanzenbach (2013) study long-run outcomes for participants in the STAR experiment, while Fredriksson, Öckert, and Oosterbeek (2013) exploit a class size rule in Sweden to estimate both short-run and long-run outcomes. These papers all find positive long-run effects of smaller classes, suggesting that the mixed effects in the literature on short-run effects are related to imperfect measurement of student skills. However, the findings for the long run are also consistent with the findings in the short run using test scores within the

[1] Summaries of the literature on the relationship between school resources and student achievement include Hanushek (1986, 2003, 2006); Krueger (2003); Webbink (2005).

[2] See Krueger and Whitmore (2001) and Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) on evidence from the STAR experiment. In contrast to the STAR experiment, field experiments on class size conducted before WW II provided little evidence in support of the hypothesis that smaller classes increase student achievement, see Rockoff (2009) for an interesting review of these early field experiments.

[3] The seminal paper by Angrist and Lavy (1999) initiated a literature exploiting class size rules in a regression discontinuity framework, Hoxby (2000) uses idiosyncratic variation in cohort size, and Wößmann and West (2006) employ a within-school across-classes strategy. While Angrist and Lavy (1999) find the expected negative effect of class size on student achievement for Israel, Hoxby (2000) and Wößmann and West (2006) find zero effects in Connecticut and for most OECD countries, respectively. In a recent paper, Denny and Oppedisano (2013) even find positive effects for the United States and the United Kingdom. They use the same empirical strategy as Wößmann and West (2006) in addition to an approach based on restrictions on higher order moments.

[4] Angrist, Battistin, and Vuri (2015) exploit a class size rule in Italy and find a strong negative relationship between test scores and class size in Southern Italy. This relationship is, however, entirely driven by manipulation of the test scores by the teachers.

same contexts.[5] Of particular interest is Fredriksson, Öckert, and Oosterbeek (2013) who find a positive short-run effect on non-cognitive ability; an outcome rarely available for researchers. These results motivate studies on long-run outcomes from contexts where the evidence indicates no class size effect on short-run outcomes.

In this paper we estimate long-run effects of class size for Norway where previous research has not been able to provide evidence of short-run gains from smaller classes in terms of student achievement.[6] We investigate whether the class size effect in lower secondary education depends on characteristics of the environment in which the schools and students operate. Leuven and Løkken (2015) explore similar data, estimating the effect of class size both in primary and lower secondary education. Their analysis, based on Leuven (2013), utilizes that some schools include grades 1 to 10, assuming that the students stayed in the same school during all school years. We find qualitatively similar effects of class size as they do and extend the analysis to investigate potential heterogeneous effects across school districts.

The findings for short-run outcomes differ substantially between the Scandinavian countries Sweden, Denmark and Norway with apparently similar educational and labor market institutions. All countries have small income differences, generous welfare state arrangements, and comprehensive public school systems seeking to equalize opportunities across families and students. Nevertheless, closer inspection reveals that important institutional differences prevail with regard to for instance school district size and teacher shortages.[7]

We first exploit the strict class size rule in Norway and match individual and school register information from 1982 through 2011 to estimate causal effects on educational attainment and income. While experimental studies are often viewed as the "gold standard" in empirical research, exploiting the class size rule in a quasi-experimental approach makes it possible to circumvent the potential Hawthorn effect that might plague experimental studies (Ehrenberg, Brewer, Gamoran, and Willms, 2001). In contrast to Fredriksson, Öckert, and Oosterbeek (2013), we are able to use register data for the whole population of schools for cohorts born 1966-1984 representing almost 1 million students and 1150

[5]See for example Krueger and Whitmore (2001) for the STAR experiment. In addition to Fredriksson, Öckert, and Oosterbeek (2013), several studies from Sweden find that increased school resources increase student performance in the short run, including Björklund, Edin, Freriksson, and Krueger (2004, ch. 4), Lindahl (2005) and Fredriksson and Öckert (2008). Browning and Heinesen (2007) and Heinesen (2010) find that lower class size in Danish compulsory education increases student performance in terms of both student test scores and educational attainment.

[6]The Norwegian studies exploiting the class size rule in short-run studies are Bonesrønning (2003), Leuven, Oosterbeek, and Rønning (2008), and Vaag Iversen and Bonesrønning (2013). They find small or zero average effects of class size. Hægeland, Raaum, and Salvanes (2012) exploit variation in school resources across school districts with different income from local taxes on hydropower plants in Norway. They find that higher resources increase student achievement.

[7]The institutional differences increased after the major reforms in Sweden in the mid-1990s. Our focus here is on institutional differences that have prevailed for several decades since several of the Swedish studies, including Fredriksson, Öckert, and Oosterbeek (2013), use data on individuals graduating compulsory education before these reforms. See Björklund, Edin, Freriksson, and Krueger (2004, ch. 4) for a description of the Swedish reforms in the 1990s and OECD (2011) and Bonesrønning (2013) for a description of recent Norwegian reforms.

schools with separate catchment areas.[8]

Secondly, information on the whole population of schools and students offers a unique possibility to use the quasi-experimental strategy to study whether the class size effect depends on characteristics of the environment in which the schools and students operate. We focus on dimensions that mirror differences in external conditions indicated by previous studies to be important for school efficiency and student performance, such as teacher quality, extent of upper secondary school choice, school district size, local fiscal constraints and labor market conditions.

We find insignificant effects of class size in grades 8-10 on educational attainment and income. While this is in contrast to the previous papers on long-run effects, it is in accordance with the findings in the short run for Norway and the long-run effect in Leuven and Løkken (2015). Moreover, we find no evidence that class size effects vary with school district characteristics.

The paper is organized as follows. In Section 2 we present arguments why the effect of resources may depend on characteristics of the external environment in which schools and students operate. Section 3 describes the institutions and the data, while the identification approach and model specification are presented in Section 4. Section 5 presents results from models estimating the causal average effect of class size on income and years of education, while Section 6 estimates interaction models investigating whether the class size effect depends on school district characteristics. Section 7 includes a discussion of the findings in relation to the present literature, and concluding comments are provided in Section 8.

2 Why might class size effects vary?

Class size may change student outcomes through a number of mechanisms affecting both student and teacher behavior. Smaller classes may be beneficial for students by reducing crowding effects through student disruption (Lazear, 2001), increasing student attention, or increasing the time teachers can use separately on each student. On the other hand, larger classes may be beneficial if a larger number of students increases the possibility that a student can find another student he/she can benefit from being in a class with, i.e., students with similar competencies, see Dobbelsteen, Levin, and Oosterbeek (2002). The literature in economics of education has also emphasized the impact of teachers, school district size and school district financing systems on student performance. In the following we discuss how these channels may affect class size effects.

[8]Fredriksson, Öckert, and Oosterbeek (2013) use data for a roughly 10 % sample of the cohorts born 1967, 1972, 1982 nd 5% sample of the cohort born 1977. In addition, to ensure exogenous catchment areas for schools, they only include school districts ("rektorsområder") with one school in their main analysis, implying that they are left with a sample of about 6000 students and 191 schools.

2.1 Teacher quality

The class size effect might depend on teacher quality as argued by educationalists (Hattie, 2005) and economists (Wößmann and West, 2006) . Hattie (2005) notes that "Without changing the teaching and ensuring rigor in the curriculum delivery then the effects of this most expensive policy is likely to be close to zero" (Hattie, 2005, p. 417). This indicates that smaller classes are only productive with high-quality teachers. Mueller (2013) uses data from the STAR experiment and finds that being assigned to a small class increases test scores when the teacher is experienced.

On the other hand, Wößmann and West (2006) conclude that "smaller classes have an observable beneficial effect on student achievement only in countries where the average capability of the teaching force appears to be low" (Wößmann and West, 2006, p. 727). This finding is supported by evidence in Altinok and Kingdon (2012), who also use an international comparable data base. They exploit subject specific class sizes in a student fixed effects strategy. We extend this line of research to an RDD framework and analyze whether the class size effect depends on teacher supply conditions.

2.2 Student incentives

The simple human capital investment model assumes that students are forward looking and make optimal educational decisions given their preferences and information on private gains and costs of education. When making educational choices, students trade off short-run costs in terms of effort in school and foregone income against future utility benefits in terms of future income.[9] Lavecchia, Liu, and Oreopoulos (2014) extend this framework to incorporate elements from behavioral economics and discuss recent empirical evidence on the relationship between student achievement and incentives provided by schools and society in the context of deviations from long-run rationality. One important element in this literature is that students are myopic and put too much weight on present effort relative to future gains. Under such circumstances external conditions affecting only short-run educational costs can be very important for future educational outcomes.

While the literature has emphasized the direct effect of student incentives, we investigate whether a gain in student achievement from increased inputs in terms of lower class size only occurs if the schools and society in general provide sufficient incentives for students to exert effort. Evidence on this issue is very limited, but Bonesrønning (2003) finds some weak evidence that class size reduction has a positive effect on test results only when teachers are able to install strong student effort incentives in terms of hard grading practices. We extend the research on student incentives to investigate whether the effect of class size is related to post-compulsory school choice systems and external labor market conditions.

[9]Examples of studies incorporating student effort in human capital investment models through educational standards is Costrell (1994), Betts (1998) and Becker and Rosen (1992).

Post-compulsory school choice

A large and still growing literature analyzes school choice as an incentive device. Although the empirical evidence is mixed, most studies find a modest positive effect of school choice and vouchers (Figlio and Hart, 2014). While school choice effects might be transmitted via a variety of mechanisms, our focus is on the effect of choice mediated by student incentives. Choice related incentives may exist in traditional public school systems. In some cases students compete for admission to different tracks within compulsory school at certain ages based on prior performance. In other cases, competition is introduced by free school choice in upper secondary education based on prior student performance. These types of competition change the incentives for students to perform well in early school years.

Koerselman (2013) finds that the change from a tracking system to comprehensive schools in England reduced test scores at early ages. Using a difference-in-differences strategy, Haraldsvik (2012) finds that the introduction of free school choice in publicly provided upper secondary education in Norway increased student performance in lower secondary education. We investigate whether the effect of class size in compulsory education is related to the extent of competition for admission into post-compulsory education.

External labor market conditions

Several studies find that student opportunity costs in terms of foregone earnings during schooling and returns to schooling are important determinants of educational attainment. Clark (2011) finds a positive effect of regional unemployment on high school enrollment in England and Wales, while Reiling and Strøm (2015) find a countercyclical pattern in high school completion in Norway. Lee (2013) finds that increased job opportunities generated by repeal of Sunday shopping restrictions in U.S. states decrease high school graduation. While these studies document the importance of job opportunities when students make educational choices after compulsory education, labor market conditions may also affect the student's allocation of time and effort during compulsory education. If class size effects depend on student incentives, the effect of class size could potentially depend systematically on labor market conditions. The fact that our data set covers a rather long time period makes it possible to investigate this issue by interacting class size with the local unemployment rate that prevails during compulsory education.

2.3 Fiscal constraints

In a traditional production function framework, more input implies higher production. Whether public sector services are produced technically efficient is, however, a widely discussed issue. In the public sector there are multiple principal-agent relationships (Dixit, 1998, 2002). Teachers and school principals might have different objectives than parents and the school district politicians. Thus, the institutional setting in which these actors operate is likely to affect the potential impact of exogenous changes in resources available for the schools. If student performance has no consequences for the decision makers in

schools, it is less likely that smaller classes would increase student performance. Instead, school principals and teachers might exploit extra resources to decrease effort, to make school days more pleasant, or to increase other types of "slack".

Some studies find evidence that decentralized decision making improves student performance (Glaeser, 1996; Barankay and Lockwood, 2007; Falch and Fischer, 2012). Hoxby (1999) argues that local funding by local property taxation can work as a discipline device on local governments and improve cost control and effort. For Norway, Borge and Rattsø (2008) provide evidence that local property taxation reduces unit costs in utility services, while Fiva and Rønning (2008) find that property taxation increases student achievement. Studies from the United States suggest that local funding increases technical efficiency in schools (Adkins and Moomaw, 2003) and student performance (Mensah, Schoderbek, and Sahay, 2013). Further, Loeb and Strunk (2007) find substantial nonlinearities in the effect of accountability policies; accountability is more effective in U.S. states with stronger local control in terms of local funding and local autonomy in hiring and spending decisions. While most of the studies so far find positive effects of local funding on efficiency and student performance, we ask whether the effect of exogenous variation in class size differ between school districts with and without access to local property tax revenue.

2.4 Interest groups

Chubb and Moe (1988) and Moe (2001, 2011) argue that teacher unions reduce the power of politicians to implement reforms and to use resources efficiently. Others argue that teacher unions may enhance efficiency by increasing teachers' job satisfaction and productivity, see Gunderson (2005) for a discussion of union voice effects in the public sector. Hoxby (1996) finds evidence that teacher unions are able to increase the teacher-student ratio, but also decrease the productivity to such an extent that student performance declines. Lovenheim (2009) finds that while unions increase teacher employment, there is no corresponding impact on student performance. Strunk and Grissom (2010) find that school districts with strong teacher unions have less flexibility in school policy than districts with weaker unions, while the evidence in Lott and Kenny (2013) indicates that students in U.S. states with strong teacher unions perform substantially worse than students in other states.

Since the large majority of teachers in Norwegian schools are members of a teacher union, it is almost impossible to study the impact of teacher unions on student performance and the interaction with class size effects. However, the impact of unions and other interest groups depends on the political setting in which they operate, i.e., by their ability to build coalitions in the government or directly affect the behavior of the decisive voter. Using survey data from Norway, Rattsø and Sørensen (2004) find that public employees prefer less public sector reform than others. Similar results are obtained by Bonesrønning (2013) who finds that school districts with a high share of public employees were less reluctant to implement a major accountability education reform in Norway in the period 2004-2006.[10]

[10] Anzia (2011) argues that members of interest groups have higher turnout in off-cycle elections than

These findings motivate studying to what extent the impact of class size differs between school districts with high and low shares of public employment.

2.5 School district size

The size of school districts varies a lot between countries. A common argument is that the competency of education governance is higher in large school districts than in small school districts. However, the evidence on scale effects in public sector production in general is mixed, and the small literature on the effect of district size on student performance is also inconclusive. For example, Driscoll, Halcoussis, and Svorny (2003) find that test scores are negatively related to district size in California. Using Danish data, Heinesen (2005) concludes that educational attainment is higher for students from larger districts, i.e. districts with population above 15,000. Berry and West (2010) exploit variation in the timing of consolidation across U.S. states and find that larger districts have some modest gains with respect to returns to education We investigate whether there is a larger return to small class size in large school districts, which are more similar to the typical school district size in Sweden and Denmark.

3 Institutions and data

3.1 Institutions

Compulsory education in Norway consists of primary schools and lower secondary schools, and ends by grade 10 the year the students turn 16 years of age.[11] Most students continue on to upper secondary education, which is divided into a three-year long academic study track and different vocational study tracks. After a major reform in 1994, vocational study tracks typically last for four years (including two years of apprenticeship training). Acceptance to an upper secondary school is based on the grades achieved in grade 10. However, all students have been guaranteed admission to upper secondary education since 1994.

There is no possibility to fail a class in compulsory education during the empirical period, implying that everyone finishes compulsory education on-time.[12] Education is comprehensive with no tracking and a common curriculum for all students. The cutoff between grades is birth at January 1.

other voters and that the policy in jurisdictions with off-cycle elections consequently are more favorable to interest groups. Consistent with this hypothesis she finds that U.S. school districts with off-cycle elections have higher teacher pay than other districts.

[11] During the empirical period, the school starting age was 7 years, but the school starting age was reduced from 7 to 6 years in 1997 such that today primary education consists of grades 1-7 (ages 6-13) and lower secondary education consists of grades 8-10 (ages 14-16). We refer to grades 8-10 as lower secondary education throughout the paper.

[12] In some cases, students do not start primary education at the expected age, which implies that they finish lower secondary education at a higher age. If a child is not considered to be mature enough, the parents together with the school and psychologists can postpone enrollment one year. In addition, some older students return to improve their grades, and immigrants are often over-aged at graduation.

Compulsory education is free of charge and is the responsibility of the municipalities. Norwegian municipalities are multipurpose institutions, providing a large number of services such as day care and care for the elderly, in addition to education.[13] In the following we refer to municipalities as school districts. There are usually several primary schools within each school district, but many small school districts only have one lower secondary school. Parental school choice between public schools for a given residence is not allowed, and private schools are quite rare and do not represent a realistic alternative to public schools. The classes could not exceed 30 students in lower secondary education during the empirical period. The class uses the same classroom for most subjects. The teachers, who are specialized in specific subjects, move between classrooms. The classes are established at the start of lower secondary education such that all classes have about the same socioeconomic composition, and it is very uncommon to change the composition of classes unless the number of classes changes.

3.2 Data

In this paper we study the cohorts born 1966-1984 who leave lower secondary education during 1982-2000. We use register data provided by Statistics Norway for all individuals leaving lower secondary education in this period. The data contain unique individual and school identifiers which allow us to combine detailed information on individuals with the school they attended.

Our two main outcome variables are years of education and income. We measure the outcomes in a given year, for which the individuals are of different age, and fully control for age effects in the empirical model. Our measure of educational attainment is years of education in 2011, measured by degrees obtained. In higher education that is bachelor degree, master degree, and PhD, with 16, 18, and 21 years of education, respectively. We use the log of average pension qualifying income for the years 2009 and 2010 as our income measure,[14] such that the youngest individuals in the sample are 25-26 years of age when income is measured.

We restrict the sample to students graduating lower secondary education the year they turn 16, which excludes 5 % of the observations. Table A1 reports the number of observations lost due to missing information on class size, the age restriction, requiring at least 10 school observations throughout the time period, and having missing information on either log income or educational attainment. We are able to use 86 % and 81 % of the population in the analysis on educational attainment and log income, respectively. The cohort leaving secondary education in 1990 has missing information on the school

[13] Spending on primary and lower secondary education accounts for about 30% of total local government spending, while spending on care for the elderly, preschool education, cultural services, infrastructure services and administration accounts for the rest.

[14] We use the pension-qualifying income as reported in the tax registry. This income measure is not top coded and includes labor income, taxable sick benefits, unemployment benefits, parental leave payments, and pensions, see Black, Devereux, and Salvanes (2013, p. 132). Information for 2011 is not available in our data.

identifier, and is thus not included in the analysis. The number of observations in the analyses is about 950,000, with cohort sizes of about 50,000 students.

The distributions of the dependent variables are presented in Figures 1 and 2. The average years of education is 14.0 with standard deviation of 2.5, while log of income has mean 12.7 with standard deviation of 0.8 (Table 1).

Figure 1: Distribution of log income conditional on cohort specific effects

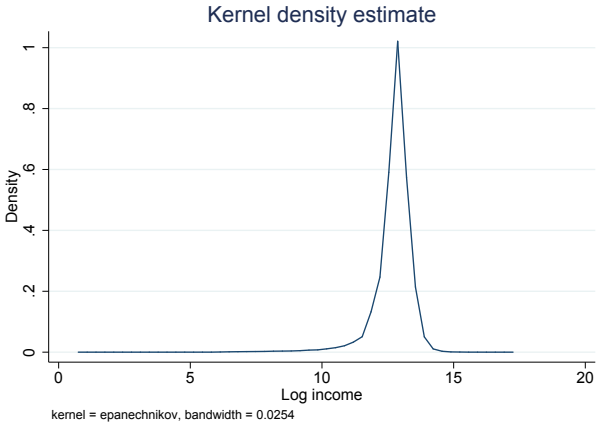
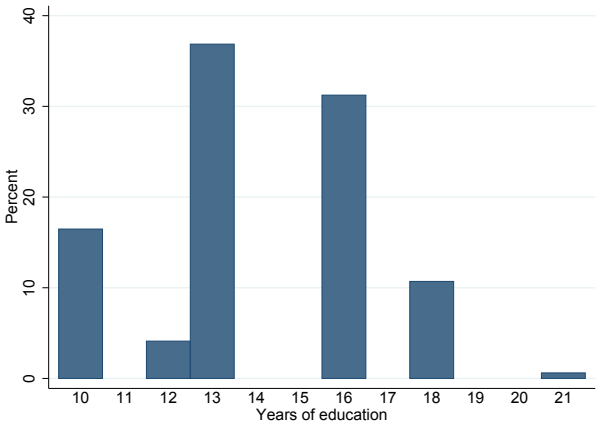


Figure 2: Distribution of years of education

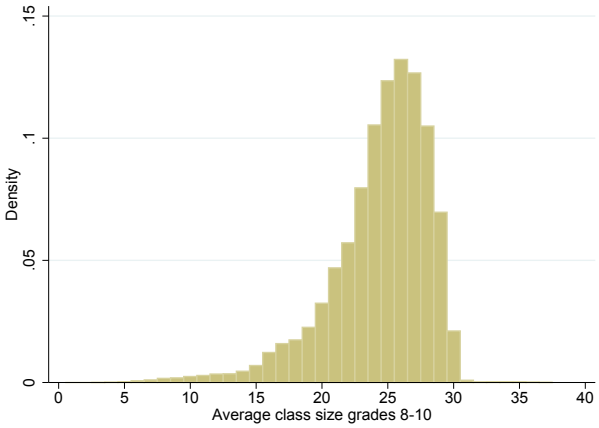


Data on the number of classes and enrollment by year and grade are obtained from a national school register administered by The Norwegian Ministry of Education. Variables are measured on October 1 of each year, which is near the beginning of the school year. The information is provided for the school rather than for the class, so we are only able to calculate the average class size for each year and grade rather than the actual class size for each class. However, a benefit of using this measure is that we do not have to worry about sorting into classes of different class sizes within schools.

Figure 3 displays the distribution of the average class size in grades 8-10 for our sample, while Table 1 provides descriptive statistics. The typical student is in a class of 23-29

students. There are extremely few observations above 30 students per class, which reflects that the class size rule is strictly followed (see also Leuven, Oosterbeek, and Rønning (2008)).

Figure 3: Average class in the empirical sample



Our individual register data contains information on gender, birth month and immigration status, as well as detailed data on educational attainment and income for all years after the individual leaves lower secondary education and up to 2011.[15] We also include information on parental education and parental employment status the year the individual turns 16 in the analysis. Descriptive statistics of the socioeconomic characteristics are presented in Table 1.[16]

4 Identification and model specification

There are several reasons why standard OLS regressions treating actual class size as an exogenous variable might yield biased estimates. For example, disruptive students with negative peer group effects might be placed in smaller classes; small remote schools with small classes might have problems in recruiting and retaining high quality teachers; student mobility might be motivated by observed class sizes; peers might correlate with class size; etc. To tackle the identification problem and estimate causal effects, one ideally want to explore only the part of variation in actual class size that is due to exogenous forces. A maximum class size rule serves this purpose.

[15]Regarding immigration status, we distinguish between first and second generation immigrants, where the former are born abroad and have both parents born abroad, while the latter are born in Norway and have both parents born abroad.

[16]Descriptive statistics on the school district characteristics used in the heterogeneity analysis are also presented in Table 1. These variables are described in Section 6 below.

Table 1: Descriptive statistics

	N	Mean	SD
A. Outcome variables			
Log of income	903,828	12.715	0.765
Years of education	952,514	13.986	2.536
B. Class size variables			
Average class size grades 8-10	952,514	24.41	3.79
Predicted class size	952,514	24.93	3.98
Enrollment grade 8	952,514	87.49	43.98
C. Socioeconomic characteristics			
Girl	952,514	0.490	0.500
Parental education: Less than high school	952,514	0.144	0.357
Parental education: High School	952,514	0.546	0.498
Parental education: Bachelor	952,514	0.202	0.401
Parental education: Master +	952,514	0.077	0.267
Parental education: Unknown	952,514	0.031	0.172
First generation immigrant	952,514	0.013	0.111
Second generation immigrant	952,514	0.006	0.076
Only mother working	952,514	0.172	0.378
Only father working	952,514	0.152	0.359
Both parents working	952,514	0.348	0.476
None of parents working	952,514	0.328	0.475
Birth month	952,514	6.342	3.335
D. School district variables			
Share of teachers with teacher certification (teacher quality)	893,546	0.960	0.039
Have school choice in upper secondary education	379,691	0.494	0.500
Unemployment rate	952,218	0.025	0.013
Have property taxation	283,322	0.379	0.485
Share of the labor force employed in the public sector	563,570	0.221	0.067
Population size	952,218	60,496	114,704
District merger: Treatment school district	952,218	0.065	0.247
District merger: Treatment school district * post-merger	952,218	0.023	0.149

Note: Descriptive statistics corresponding to the estimation sample for years of education.

4.1 The class size rule

During the time period we study, a national rule was in place saying that class size could not surpass 30 students in lower secondary education. The class size rule creates exogenous variation in predicted class size depending on the number of students enrolled in a school.

Since learning is cumulative, we estimate the effect of average class size during lower secondary education (grades 8 to 10) and not the class size in one specific school year. Each student is matched to their lower secondary school at graduation, and we use information from this school also for the two previous school years to calculate average class size.[17] For each grade level the data contain the number of classes and the number students enrolled.

We follow Leuven, Oosterbeek, and Rønning (2008) and use predicted class size based on enrollment in grade 8, two years prior to graduation, as the instrument in the analysis in order to avoid biased estimates due to possible endogenous mobility of students across schools during the years in lower secondary education. The instrument is given by

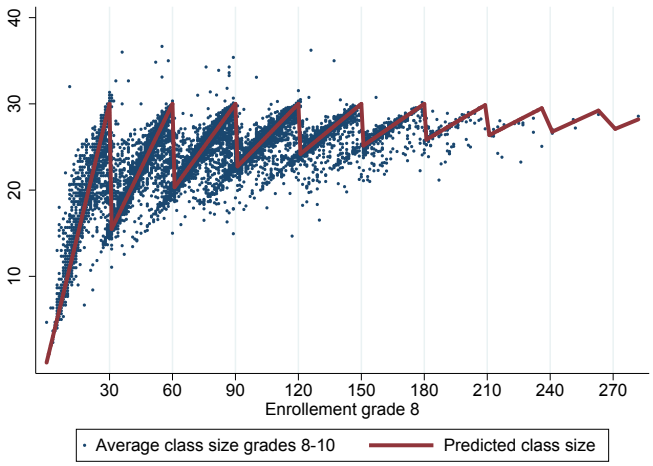
$$CS_{t-2}^{rule} = \frac{E_{t-2}}{int(1 + (E_{t-2} - 1)/CS^{max}} \tag{1}$$

where E_{t-2} is enrollment in grade 8 and CS^{max} is the maximum class size according to the rule. Using this formula, the strict maximum class size rule predicts a class size of 30 when 30 students are enrolled and a class size of 15.5 when 31 students are enrolled. Such a kink appears at each multiple of 30 and creates a nonmonotonic relationship between enrollment and predicted class size. We follow Angrist and Lavy (1999) in instrumenting actual class size by predicted class size defined in equation (1), while controlling flexibly for enrollment.

Figure 4 plots the class size rule calculated by equation (1) and average actual class size for grades 8-10 against enrollment in grade 8. Average class size closely tracks the class size rule for all enrollment levels.

[17] The average class size is calculated using information on grades 10, 9 and 8 in year t, t-1 and t-2, respectively, i.e., when the student was enrolled in the relevant grades.

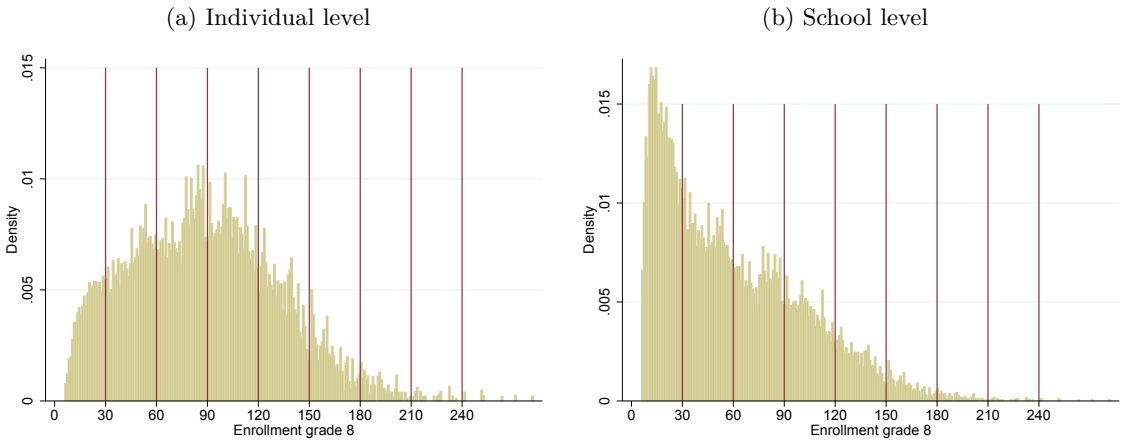
Figure 4: The first stage



One possible threat to the validity of the instrument is manipulation of enrollment around the thresholds. Urquiola and Verhoogen (2009) find this to be the case in Chile. Fredriksson, Öckert, and Oosterbeek (2013) also find that sorting take place within school districts because "it is likely that school catchment areas are adjusted" (Fredriksson, Öckert, and Oosterbeek, 2013, p. 254). Thus, their analysis includes only school districts with one school.

In Norway, it has been uncommon to adjust school catchment areas. Panel A in Figure 5 plots the distribution of enrollment in grade 8, where the vertical lines represent the class size thresholds. There is no evidence of manipulation of the enrollment. The density of observations just below and above the thresholds is similar. In fact, the enrollment is higher just above the threshold in 5 out of the 8 class size thresholds in the data. In addition, the figure shows that it is mainly the thresholds at enrollment of 30, 60, 90, and 120 students that will contribute to the identification of the class size effect. While the density in Panel A in Figure 5 is presented at the individual level, the identification is at the school level. Panel B uses the school as the observational unit, and shows that few schools have enrollment above 150 students in grade 8. Most schools have enrollment around the first threshold, for which there is the largest difference in class size across the threshold.

Figure 5: Distribution of enrollment in grade 8 in the empirical sample



A more direct way to assess whether the instrument is valid is to examine whether socioeconomic characteristics are equal across observations above and below the class size thresholds. Table 2 tests the balancing of the covariates both individually and jointly.

The first two columns in Table 2 show that the socioeconomic characteristics are strong predictors of income and education as expected. The correlation with parental education is particularly strong. Column (3) presents results for a regression on the class size rule, using the control function for enrollment described below. None of the socioeconomic characteristics are significant at the 5% level, and the test for joint significance has a p-value of 0.08. Column (4) presents p-values for individual correlations, which are significant at the 5% level for two measures of parental education. Overall, however, the socioeconomic characteristics in the data are reasonably unrelated to the class size rule.

4.2 Model specification

We present results from two approaches to the regression discontinuity design. The first approach uses all information available, and includes a flexible control for the effect of cohort size at the school. The second approach discards observations away from the thresholds and uses a simpler specification for cohort size, see for instance Lee and Lemieux (2010) and Gelman and Imbens (2014) for discussions of these approaches. We denote the former a "global" approach and the latter a "local" approach. In both approaches it is important to control for age effects because income and education are measured in a specific year, and thus at different ages. Since the analysis only includes individuals graduating lower secondary education at age 16, including cohort fixed effects is identical to including age fixed effects in our application.

Both approaches imply that we estimate variants of the following model

$$y_{ist} = \alpha \overline{CS}_{st} + f(E_{st-2}) + \beta X_i + \delta_t + \varepsilon_{ist} \tag{2}$$

where y_{ist} denotes the outcome for individual i graduating from school s in year t and

Table 2: Balancing sample

	Log income	Years of education	Predicted class size	p-value
	(1)	(2)	(3)	(4)
Girl	-0.336***	0.545***	-0.0013	0.768
	(0.0033)	(0.0081)	(0.0046)	
Parental education: High School	0.114***	1.049***	-0.0127	0.010
	(0.0025)	(0.0098)	(0.0080)	
Parental education: Bachelor	0.163***	2.391***	0.0039	0.017
	(0.0033)	(0.0121)	(0.0090)	
Parental education: Master +	0.176***	3.349***	-0.0069	0.986
	(0.0051)	(0.0158)	(0.0112)	
Parental education: Unknown	0.009	0.750***	-0.0002	0.685
	(0.0062)	(0.0250)	(0.0158)	
First generation immigrant	-0.058***	-0.032	0.0012	0.973
	(0.0104)	(0.0366)	(0.0339)	
Second generation immigrant	0.030**	0.455***	0.0955	0.157
	(0.0147)	(0.0438)	(0.0682)	
Only mother working	0.042***	0.129***	-0.0035	0.185
	(0.0026)	(0.0092)	(0.0069)	
Only father working	0.039***	0.054***	-0.0014	0.400
	(0.0026)	(0.0094)	(0.0071)	
Both parents working	0.108***	0.458***	0.0099	0.081
	(0.0024)	(0.0092)	(0.0069)	
Birth month	0.0007***	0.006***	0.0001	0.876
	(0.0002)	(0.0007)	(0.0007)	
Observations	903,828	952,514	952,514	
R-squared	0.107	0.151	0.305	
Number of grunn_id	1,156	1,156	1,156	
p-value of F-test	0	0	0.0809	

Note. Columns (1)-(3) report results of OLS regressions on the variables listed in the rows, where predicted class size is our class size instrument. These regressions also include the following control variables: fixed effects for enrollment segment, enrollment to the fourth polynomial, and time/age fixed effects. Independent variables are pre-determined parent and student characteristics. The p-value reported at the bottom of columns (1)-(3) is for an F-test of the joint significance of the variables listed in the table. Each row of column (4) reports a p-value from separate OLS regressions of the pre-determined variable (listed in the corresponding row) on the instrument, and the same set of control variables as in columns (1)-(3). Estimates in column (3) and (4) correspond to the sample used for educational attainment. The p-value is for a t-test of the significance of the class size instrument. Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001. Standard errors are clustered at the school level.

$\overline{CS}$ is the predicted average class size for grades 8-10. In addition, the model includes a flexible functional form of enrollment E in grade 8, individual characteristics, X_i, and cohort fixed effects (δ_t). The error term (ε_{ist}) is clustered at the school level. The first stage is simply

$$\overline{CS}_{st} = \alpha' CS^{rule}_{st-2} + f(E_{st-2}) + \beta' X_i + \delta'_t + \varepsilon'_{ist} \tag{3}$$

When using the "global" approach, a flexible modelling of enrollment effects in terms of the function $f(E_{st-2})$ is necessary in order to avoid that the discontinuity generated by the class size rule is confounded with a possible nonlinear relationship between the outcome variable and enrollment. Define the thresholds for the class size rule in grade 8 as $\tilde{E}_{st-2} = \{30, 60, 90, ...270\}$, and the segments of the class size rule as $S_{st-2} = I(\tilde{E}_{st} \pm 15)$. The following specification for the global approach seems to capture both the underlying functional form and to provide reasonable precision of the estimates

$$f(E_{st-2}) = \alpha_1 E_{st-2} + \alpha_2 E^2_{st-2} + \alpha_3 E^3_{st-2} + \alpha_4 E^4_{st-2} + \alpha_5 S_{st-2} + \delta_s \tag{4}$$

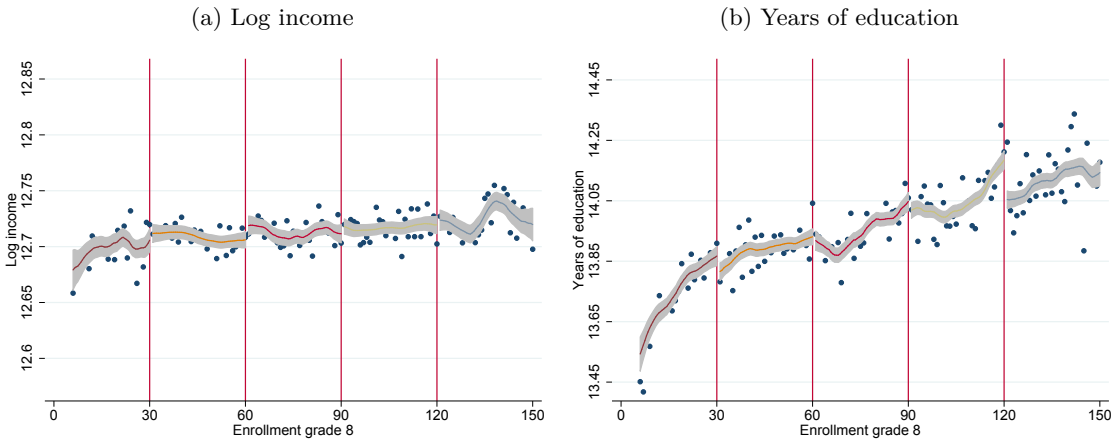
where δ is school fixed effects.

The global approach essentially uses a bandwidth of ± 15 students. The local approach uses a substantially smaller bandwidth. In the case with the smallest possible bandwidth and only one discontinuity, $\left[\tilde{E}_{st-2}, \tilde{E}_{st-2} + 1\right]$, it is not possible to control for enrollment. The identifying assumption is that the outcome at these two enrollment levels would be equal in the absence of the discontinuity. Since we have several threshold levels in the data, we estimate local effects with the following model specification of enrollment:

$$f(E_{st-2}) = \alpha'_1 E_{st-2} + \alpha'_5 S_{st-2} \tag{5}$$

Figure 6 present average values of the outcomes for different levels of enrollment and shows that the outcomes are positively related to enrollment. Since average class size is higher in larger schools than in small schools, this implies that class size and the outcomes are positively related, in contrast to the hypothesized class size effect. The local polynomial regressions presented in the figure do not indicate any systematic changes in the outcomes related to the thresholds. For income, there seems to be a difference for the threshold of 60 students in the expected direction. For educational attainment, there seems to be differences both for the thresholds 30 and 120 students, but in the opposite direction of what is expected.

Figure 6: Local polynomial regressions



Note: Local polynomial regressions of enrollment in grade 8 on outcome variables for each segment. Log income and educational attainment are conditional on cohort specific effects. The markers indicate average outcome for each enrollment value. The y-axis is the mean value of the outcome variable +/- 0.2 standard deviations.

5 Average class size effects

For the global approach, in which all observations in the data are used, the results for different model specifications are presented in columns (1) - (8) in Table 3. Column (1) presents a simple OLS regression with cohort fixed effects and a linear enrollment control. With this specification, children in larger classes have higher income (t-value of 0.72) and complete more years of schooling (t-value of 5.60) than children in smaller classes, contrary to the expectations. However, when average class size is instrumented in this very simplistic model formulation (column (2)), the class size effect on income gets the expected sign, but is still insignificant. Predicted class size is a strong instrument. The F-value for the first stage is almost 5,000.

Columns (3)-(8) include various specifications of the enrollment control function. Regarding income, the point estimate is negative and clearly insignificant in all specifications. The result for educational attainment is more sensitive to the specification of the enrollment control function. The effect is positive and significant at 5% level in the models only including segment fixed effects (column 3) and enrollment to the fourth polynomial (column 4). When school fixed effects are introduced (column 5), the effect drops and becomes insignificant.

Column (6) additionally includes socioeconomic characteristics. This does not affect the class size effect, as expected from the balancing tests in Table 2. In column (7) and (8), enrollment is interacted with segment fixed effects. While the interaction is linearly in column (7), column (8) also includes interaction with enrollment up to the fourth polynomial. Although the strength of the instrument declines as the enrollment control function becomes more flexible, the F-value for the first stage is above 900 in each specification.

Table 3: Average class size effect and specification analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Log income										
Average class size grades 8-10	0.00035	-0.00033	-0.00077	-0.00087	-0.00029	-0.00048	-0.00046	-0.00099	-0.00002	0.00025
	(0.0005)	(0.0005)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0011)	(0.0021)	(0.0030)
F-value first stage	-	4,843	2,391	2,327	1,975	1,976	1,864	911.4	295.9	146.1
Observations	903,828	903,828	903,828	903,828	903,828	903,828	903,828	903,828	170,604	170,604
B. Years of education										
Average class size grades 8-10	0.0158**	0.0117**	0.0118**	0.0106**	0.0017	0.0007	0.0003	-0.0001	-0.0012	0.0089
	(0.0028)	(0.0032)	(0.0038)	(0.0038)	(0.0023)	(0.0021)	(0.0021)	(0.0034)	(0.0069)	(0.0103)
F-value first stage	-	4,827	2,386	2,322	1,974	1,975	1,863	909.0	294.7	145.4
Observations	952,514	952,514	952,514	952,514	952,514	952,514	952,514	952,514	179,799	179,799
Estimation method	OLS	IV	IV	IV	IV	IV	IV	IV	IV	IV
Enrollment controls	Linear	Linear	Linear and seg. FE	Polynomial and seg. FE	Polynomial and seg. FE	Polynomial and seg. FE	Linear and int. with seg. FE	Polynomial and int. with seg. FE	Linear with seg. FE	Linear and int. with seg. FE
Time/age fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School fixed effects	No	No	No	No	Yes	Yes	Yes	Yes	No	No
Socioeconomic characteristics	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Sample	All	All	All	All	All	All	All	All	$\tilde{E}_{st-2} \pm 3$ students	$\tilde{E}_{st-2} \pm 3$ students

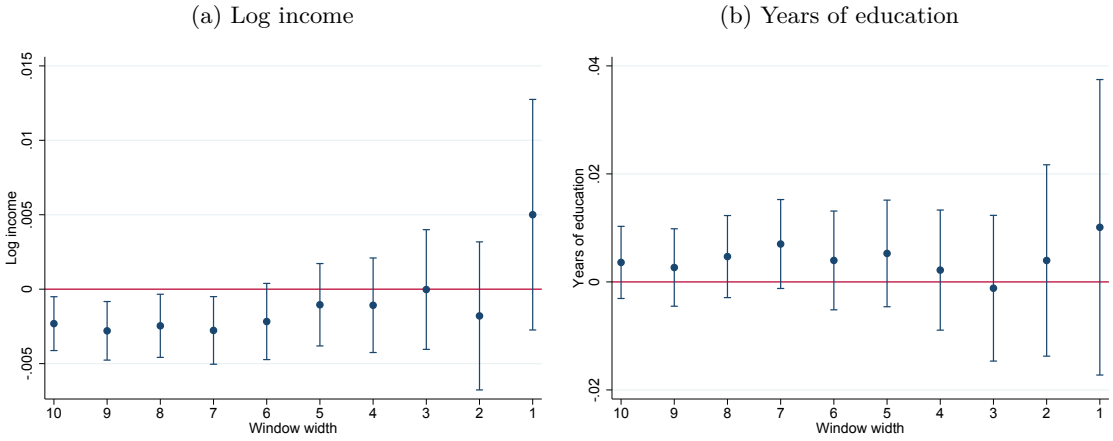
Note: Standard errors clustered at the school level in parentheses, * p<0.05 and ** p<0.01. Socioeconomic characteristics include birth month, gender, immigration status, parental education, and parental employment status. Full model results for columns (6) and (9) are presented in Appendix Table A3. When describing enrollment controls, seg. is segment and int. is interacted.

Column (6) is the model specification in equations (2) – (4) above. Taken at face value, the 95% confidence interval of reduced class size of 10 students is [-0.018, 0.008] log points for income and [-0.029, 0.052] for years of education. Both intervals are very narrow. We can rule out even very small effects of class size.

The full results for the models in column (6) are presented in Appendix Table A2 columns (1) and (3). The effects of socioeconomic characteristics are as expected. Females have longer education than males, but lower income. In addition, Table A2 shows results for the first stage. The first stage coefficient is 0.56, which is very close to the result in Leuven, Oosterbeek, and Rønning (2008) despite that they only include students graduating lower secondary education in 2002 and 2003.

Figure 7 presents estimates for the local approach with 95% confidence intervals, shrinking the bandwidth from ±10 students to ±1 student. In the latter case, only observations just below and just above the thresholds are included (30 and 31 students, 60 and 61 students, etc.). The model formulation is equal to equations (2) and (5) above, and the results for bandwidth of ±3 students are presented in column (9) in Table 3.[18]

Figure 7: Effect of class size with 95% confidence interval when reducing bandwidth from 10 to 1.



For educational attainment, the estimated effects are insignificant for all bandwidths, and the point estimate is positive in all cases except one. Increased years of education for larger classes is in contrast with the intuitive hypothesis. For income, the point estimate is negative for all bandwidths except the most narrow. For large bandwidths, the effect is close to -0.002 and statistically significant at conventional levels. This is a stronger effect than for the global approach, but the enrollment control function is rather simplistic in these models because it is specified for a model with a narrower bandwidth. For bandwidths of ±6 students or smaller, the estimated effect is smaller and insignificant. Column (10) in Table 3 presents results for a model with a more flexible enrollment control function, including enrollment interacted with the segment fixed effects, for a bandwidth

[18]For a full specification of the models, se columns (2) and (4) in Appendix Table A2.

of ±3 students. This changes the sign of the effect on both income and educational attainment, but the effects are still clearly insignificant. The strength of the instrument is reasonable also in this case with F-value for the first stage above 100.

To shed some light on what can be driving the insignificant results, Figure 8 presents cohort specific estimates using the model specification in column (6) in Table 3. The oldest cohort is born in 1966, graduated from lower secondary education in 1982, and years of education is measured at age 45 while income is measured as the average income at age 43 and 44. The estimate is not significant at the 5% level for any cohort and any outcome. For income, the point estimate is positive for four of the 19 cohorts, while for educational attainment, the estimate is positive for 12 cohorts.[19]

Figure 8: Cohort specific estimates using the global approach with 95% confidence intervals.

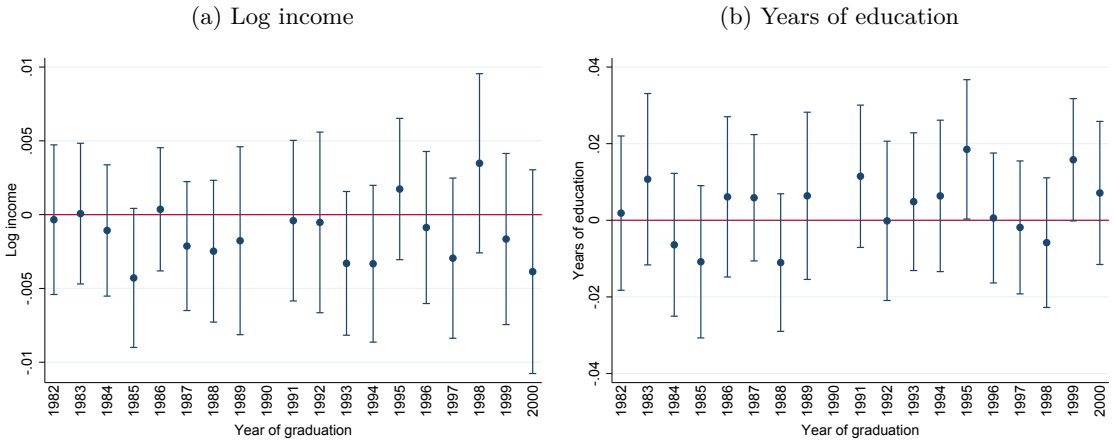
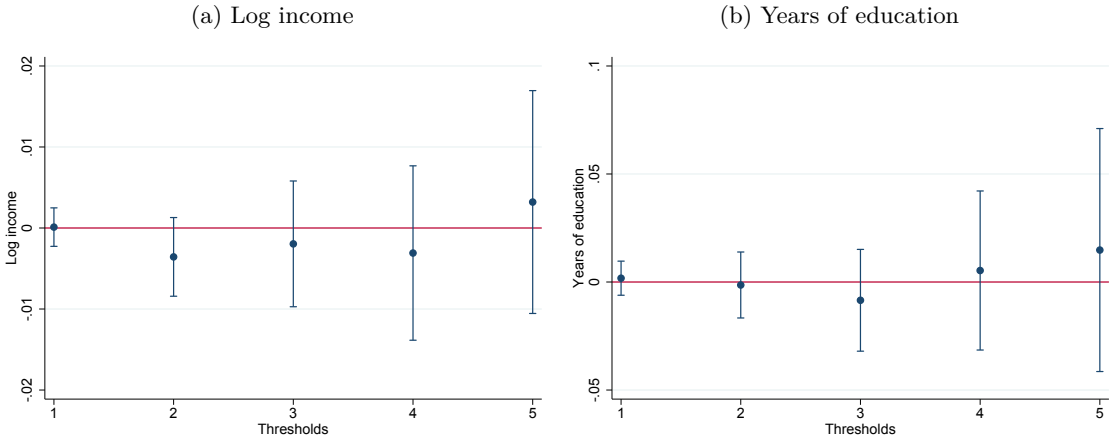


Figure 9 presents separate analysis for the different thresholds. The regressions are equivalent to column (9) in Table 3, with the segment fixed effects absorbed by the constant term. The regression denoted threshold 5 includes all thresholds from 5 (150 students) and upwards. As expected, confidence intervals at the 95% level rise with each threshold. In all cases, the effect of class size is insignificant at 5 percent level and close to zero.[20]

[19]Log income has a wide distribution, see Figure 1. However, this does not drive the results. In regressions including only observations with log of income between 10 and 15 (reduces the sample by 1.4%), the estimate for average class size is -0.00011 (0.0005) using the global approach specification in column (6) and -0.00098 (0.0016) using the local approach in column (9).

[20]We have also run regressions using a binary variable for whether the student achieves a degree from higher education (completes more than 13 years of schooling) as an outcome variable. The effect is insignificant also for this measure of educational attainment. The estimate of average class size is 0.0006 (0.0004) using the global approach specification in column (6) in Table 3 and -0.0003 (0.0014) using the local approach in column (9).

Figure 9: Effect of class size with 95% confidence interval when running separate regressions for each threshold.



Note: Threshold 5 includes all thresholds from 5 and up.

One common argument for smaller classes is that it can improve the possibility to support students most in need of learning support. The evidence from for instance the STAR experiment suggests that students with a disadvantaged background benefit the most from smaller classes (Dynarski, Hyman, and Schanzenbach, 2013), which suggests that smaller classes have the potential to reduce the variation in student outcomes. We investigate this issue in Table 4 where we use data collapsed to school-by-year observations, and the standard deviation in the outcomes are the dependent variables.[21] Columns (1) and (2) presents results for the global approach, while columns (3) and (4) presents results for the local approach. The measures of the socioeconomic composition at the schools included in columns (2) and (4) are simply the average values over the relevant students for the individual characteristics presented above.

The effects of smaller classes in column (1) in Table 4 are positive as expected. Reduced class size of 10 students significantly decreases the variation in log income by 0.03, which is 12% of a standard deviation. The effect on the variation in years of education is about 6% of a standard deviation, but insignificant. However, including measures of the socioeconomic composition at the school reduces the class size effect considerably.[22] In addition, the effects estimated by the local approach are negative and insignificant. Overall, it does not seem like smaller classes reduces the variation in student outcomes.

6 Heterogeneous class size effects

In this section we investigate whether the class size effect depends on the external environment in which schools and students operate as discussed in Section 2 above. We focus on measures of teacher quality, fiscal constraints facing school districts, variables affecting

[21] The average values (standard deviation) for the dependent variables are 0.69 (0.26) and 2.42 (0.32) for the standard deviation in log income and years of education, respectively.

[22] The effect disappears when controlling for parental education.

Table 4: Effect of class size on variation in outcomes, school level analysis

	(1)	(2)	(3)	(4)
A. Dependent variable: Standard deviation in Log income				
Average class size grades 8-10	0.00322**	0.00158	-0.00322	-0.00538
	(0.0012)	(0.0011)	(0.0037)	(0.0035)
F-value first stage	2,620	2,605	215,1	223.0
Observations	16,731	16,731	2,713	2,713
B. Dependent variable: Standard deviation in Years of education				
Average class size grades 8-10	0.00204	-0.000280	-0.00126	-0.00538
	(0.0014)	(0.0014)	(0.0041)	(0.0040)
F-value first stage	2,623	2,607	215,1	223.0
Observations	16,734	16,734	2,713	2,713
Enrollment controls	Pol. and seg. FE	Pol. and seg. FE	Linear and seg. FE	Linear and seg. FE
School fixed effects	Yes	Yes	No	No
Time/age fixed effects	Yes	Yes	Yes	Yes
Socioeconomic composition at school	No	Yes	No	Yes
Subsample +/- 3 students	No	No	Yes	Yes

Note: Standard errors clustered at the school level in parentheses, * p<0.05 and ** p<0.01. Socioeconomic composition at school is measured as average values of the socioeconomic characteristics included in Table 2. When describing enrollment controls, seg. is segment and pol. is polynomial.

student effort incentives, variables affecting interest group pressure, and school district size. All variables are measured at the school district level. The small average treatment effect of class size in the long run might hide differences across school districts, and specific characteristics in some Norwegian school districts might explain the different average results compared to Dynarski, Hyman, and Schanzenbach (2013), Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011) and Fredriksson, Öckert, and Oosterbeek (2013).

For each school district characteristic Z of interest, we estimate the following model

$$y_{isdt} = \alpha \overline{CS}_{sdt} + \gamma Z_{dt} + \phi \overline{CS}_{sdt} \times Z_{dt} + f(E_{sdt-2}) + g(E_{sdt-2}) \times Z_{dt} + \beta X_i + \delta_t + \varepsilon_{isdt} \quad (6)$$

where subscript d indicates school district. This is equivalent to estimating equations (2) and (3), adding Z and the interaction terms with average class size and the control function for enrollment. The control functions $f(.)$ and $g(.)$ include the same elements as above. $\overline{CS}_{sdt}$ and $\overline{CS}_{sdt} \times Z_{dt}$ are instrumented using the class size rule and its interaction with Z. Since we use average class size during grades 8-10 in the analysis, we measure the school district characteristics by the average value during the same time period. In order to facilitate interpretation, the interaction variables are standardized to have mean zero and standard deviation equal to unity, except when indicated. The level effect of Z is not reported since the interaction term with $g(.)$ is included in the model.

6.1 Teacher quality

The evidence in the literature on the relationship between a class size effect and teacher quality is mixed. One empirical challenge is that teacher quality is not directly observed. Our approach is that teacher quality is related to the attractiveness of the school. According to the Norwegian school law, schools can only employ persons without a teaching certification if no certified teacher apply to a vacant teacher position, and non-certified teachers can only be employed for up to one school year. Teacher shortages measured by non-certified teachers thus reflect the state of the teacher labor market in a particular year. If the use of non-certified teachers increases, it reflects low interest for vacant positions, lack of options in the schools' hiring processes, and thus low teacher quality. The share of certified teachers is thus a reasonable indicator of teacher quality, and is previously used by Bonesrønning, Falch, and Strøm (2005) and Falch, Johansen, and Strøm (2009).

The first part of Table 5 presents the results.[23] Columns (1) and (3) use the global approach, while columns (2) and (4) use the local approach. The level effects of average class size are close to the findings in Table 3 as expected since the measure of teacher quality is standardized.[24] The joint strength of the instruments is tested by the Kleibergen-Paap F-statistic, and the test value above 100 implies that the instruments are not weak.

The interaction effect with our measure of teacher quality is negative or close to zero. The sign of the coefficient indicates that class size might have the expected negative effect when teacher quality is high. The best teachers might to be able to exploit the possibilities inherent in small classes. For the income-equation using the local approach, the interaction effect is significant at 5% level. The results imply that decreasing class size by 10 students in school districts with teacher quality 2 standard deviations above the average, increases the income by 0.054 log points; about 7% of a standard deviation in income.

6.2 Student incentives

Without student incentives, more resources can hardly improve student achievement. We investigate the effect of two different student incentives that are external to the school district authorities. First, upper secondary education is non-compulsory and is the responsibility of the 19 counties. Some counties have free school choice, while other counties use school catchment areas. With free school choice, the students rank schools in their applications, and admission to oversubscribed schools is solely based on grade point average from lower secondary education (GPA).[25] Thus, there are stronger incentives for

[23] Data for our measure of teacher quality is available from 1981. However, since we use 3 year averages in the estimations, the samples used in the analyses are from 1983 and onwards.

[24] Notice that since we have rescaled the variable for teacher quality to have mean zero, there are only two reasons why the level effect of class size could differ from the similar model in Table 3. First, the model includes an additional variable (teacher quality), and second, the sample size is about 5 % smaller. If we re-estimate the corresponding models in Table 3 using the same sample as in Table 5, we get the same coefficients on class size.

[25] A closer description of one system of free school choice is given in Machin and Salvanes (2016). They study the effect on house prices of increased school choice from 1997 in the Oslo county.

Table 5: Heterogeneous effects of class size

	Log income		Years of education	
	(1)	(2)	(3)	(4)
A. Teacher quality				
Interaction effect with class size	0.00005	-0.00270*	-0.0022	0.0042
	(0.0006)	(0.0016)	(0.0018)	(0.0050)
Average class size grades 8-10	-0.00041	-0.00028	0.00003	0.0018
	(0.0007)	(0.0023)	(0.0022)	(0.0076)
F-value, first stage	598.0	100.00	597.1	101.0
Observations	849,163	159,830	893,546	168,182
B. School choice upper secondary education				
Interaction effect with class size	0.00048	-0.00668	0.0089	-0.0186
	(0.0022)	(0.0078)	(0.0067)	(0.0257)
Average class size grades 8-10	-0.00008	0.00244	-0.0016	-0.0031
	(0.0013)	(0.0041)	(0.0039)	(0.0143)
F-value, first stage	204.3	19.62	203.4	19.82
Observations	364,670	69,101	379,619	71,876
C. Local unemployment rate				
Interaction effect with class size	-0.00016	-0.00244	-0.0019	-0.0119
	(0.0007)	(0.0022)	(0.0021)	(0.0082)
Average class size grades 8-10	-0.00031	-0.00054	0.00075	0.0008
	(0.0007)	(0.0020)	(0.0021)	(0.0068)
F-value, first stage	287.9	17.26	292.0	17.25
Observations	903,572	170,601	952,218	179,796
D. Property tax				
Interaction effect with class size	0.00122	-0.01050	-0.0046	0.0269
	(0.0028)	(0.0089)	(0.0087)	(0.0317)
Average class size grades 8-10	-0.00109	0.00851*	0.0033	-0.0194
	(0.0015)	(0.0052)	(0.0041)	(0.0155)
F-value, first stage	128.4	13.71	127.7	13.54
Observations	272,724	52,087	283,322	54,053
E. Share of public sector employment				
Interaction effect with class size	0.00054	-0.00099	-0.0032	0.0121
	(0.0008)	(0.0030)	(0.0026)	(0.0096)
Average class size grades 8-10	-0.00012	-0.00061	0.0016	0.0012
	(0.0009)	(0.0029)	(0.0028)	(0.0094)
F-value, first stage	375.8	14.03	370.6	13.75
Observations	539,693	101,333	563,569	105,745
F. School district size; population				
Interaction with class size	0.00125	-0.00184	-0.0006	-0.0017
	(0.0009)	(0.0039)	(0.0022)	(0.0103)
Average class size grades 8-10	0.00002	0.00012	0.0001	0.0007
	(0.0007)	(0.0023)	(0.0021)	(0.0076)
F-value, first stage	61.83	8.062	60.72	8.152
Observations	903,572	170,601	952,218	179,796
G. School district size; merger				
Interaction with class size (treatment school district * Post-merger * average class size)	0.00780	0.02860	-0.0136	0.0780
	(0.0082)	(0.0278)	(0.0222)	(0.0676)
Average class size grades 8-10	-0.00024	-0.00011	0.0004	-0.0018
	(0.0007)	(0.0021)	(0.0021)	(0.0070)
Average population in the school district during grades 8-10, standardized	-0.246***	-0.0123***	0.456***	-0.0325
	(0.0473)	(0.0030)	(0.0828)	(0.0217)
F-value, first stage	9.456	1.687	9.562	1.659
Observations	903,572	170,601	952,218	179,796
Enrollment controls	Pol. and seg. FE	Linear and seg. FE	Pol. and seg. FE	Linear and seg. FE
School FE	Yes	No	Yes	No
Subsample +/- 3 students	No	Yes	No	Yes

Note. Standard errors clustered at the school level in parentheses, * p<0.05 and ** p<0.01. The model specifications are equal to the model specifications in column (6) and (9) in Table 3, except as indicated. Instruments for average class size in grades 8-10 and the interaction effect with class size is the class size rule in grade 8 and the interaction with the class size rule in grade 8. When describing enrollment controls, seg. is segment and pol. is polynomial.

study effort in lower secondary education in some counties than in others.[26] We use the classification developed by Haraldsvik (2003),[27] previously exploited by Falch and Naper (2013). Indeed, Haraldsvik (2012) finds that school choice in upper secondary education in Norway increases student achievement in lower secondary education. Our hypothesis is that since school choice increases student incentives, the effect of class size is larger than without school choice. The results are presented in the second part of Table 5. The effect of the interaction between class size and the dummy variable for free school choice is negative as expected when using the local approach, but insignificant at conventional levels in all models. Taken at face value, the point estimate in the case of school choice of a reduction in class size of 10 students is 0.067 log points on income and 0.19 years of education.

Our second measure of student incentives is the unemployment rate in the school district. The interaction effects are negative as expected, but small and insignificant. Again the estimated class size effect is largest on income in the case with local identification, and of comparable size as in the model for teacher quality. But taken together, the results indicate that student incentives does not have a robust impact on how efficient schools use their resources.

6.3 Fiscal constraints

Local funding by local property taxation can work as a discipline device on local governments and lead to better cost control (Glaeser, 1996; Hoxby, 1999). In Norway, some school districts have property taxes while others do not. We exploit this variation in order to investigate whether class size has the expected effect with stronger fiscal constraints, i.e., there is a stronger incentive for cost control and effort.

Local governments decide both on the valuation of houses, the tax-free allowance, and the tax rate, but data on these properties of the local tax systems are not available. In our analysis we follow Borge and Rattsø (2008) and use an indicator for whether the school district has property taxation or not, for which comparable data are available in the period 1997-1999. Introduction or abolishing of property taxation are political decisions with strong local interest, and does not happen often. The share of school districts with property taxation is 14.0 – 15.6 percent in this period, and is most common in the large school districts. Since we use three-year averages of the variables in the analyses, we extrapolate the information on property taxation in both ends, assuming that the values are the same in 1995 and 1996 as for 1997, and the same in 2000 as for 1999. The

[26] In addition, the students have to rank three different study tracks in their application to upper secondary education. They have a legal right to be enrolled into one of these three tracks, but whether they are enrolled in the first, second, or third preferred track depends on their GPA.

[27] Haraldsvik (2003) distinguishes between school districts where the students have (i) free school choice between at least five schools or (ii) with some limitations, (iii) free school choice but between less than five schools, (iv) no choice at all, and (v) some marginal school choice. We classify the former three school districts as free school choice and the two latter school districts as without school choice. School districts were in 2003 asked about their school choice rules for the past 10 years. The regression sample is therefore from 1993 and onwards.

estimation period is therefore 1995-2000.

The results in Table 5 are again insignificant at the conventional level, and the sign of the interaction effect varies across the model specifications. The class size effect seems to be unrelated to local fiscal constraints.

6.4 Interest groups

Interest groups prefer increased resource use and reduced pressure on efficiency. As discussed in Section 2, there is some evidence in the literature indicating that public sector employees are more prone to interest groups than others. We use the share of public sector employment as an indicator of interest group influence, including employees both in local governments and the central government, and test the hypothesis that the class size has a larger negative effect when this share is low.[28]

Table 5 shows that also this interaction term is insignificantly related to the class size effect. The point estimates are small, and the sign of the effects varies across the specifications.

6.5 School district size

Are the resources used more efficiently in school districts with presumably more competent management of the schools? In the Norwegian setting it is usually argued that small school districts have challenges recruiting quality leadership and implementing efficient governance systems, which also was the main argument for the major school district consolidation in Denmark in 2007. There is a positive relationship between student achievement and school district size in the Norwegian data.

We investigate the interaction between the class size effect and school district size in two different ways. Firstly, we include interaction effects with the number of inhabitants in the school district. In this case the interaction effect is mainly negative as expected, but clearly insignificant. The F-value of the test of weak instruments is smaller in these models than in the models above, most likely because the schools are larger in the cities. Population size and predicted class size are positively correlated.

In general, the interaction effect with class size in this case might reflect unobserved characteristics of the school district. In addition, since the model using the whole sample includes school fixed effects and population changes only to a small extent from one year to another, little variation in school district size is used for identification in this case. Our second approach exploits that some school districts have merged during the empirical period, while the schools' catchment areas did not change.

We combine a difference-in-differences approach with regard to school districts merging and the regression discontinuity approach with regard to class size. The model includes an indicator variable for whether or not the school districts ever experiences a merger

[28] Information on the share of public sector employment in the school district is available from 1984, which implies that the regression samples are from 1986 and onwards.

($Treat$) and an indicator variable for the period after the merger in the treated school district ($Post$), in addition to the population size (Pop).

$$\begin{aligned} y_{isdt} &= \alpha\overline{CS}_{sdt} + \gamma_1 Treat_d + \gamma_2 Treat_d \times Post_t + \phi\overline{CS}_{sdt} \times Treat_d \times Post_t \\ &\quad + f(E_{sdt-2}) + g(E_{sdt-2}) \times Treat_d \times Post_t + \beta_1 Pop_{dt} + \beta X_i + \delta_t + \varepsilon_{isdt} \end{aligned} \tag{7}$$

The term $\gamma_2 + g(E_{sdt-2})$ is the difference-in-differences estimator. Both terms including class size are instrumented in the same way as above.

Results are reported towards the end of Table 5. The results for the local approach can hardly be interpreted in this case because the instruments are weak. For the global approach, the interaction effects are relatively large, but insignificant and with opposite sign for income and education.

7 Discussion

Contrary to the results for the United States, Sweden and Denmark, we find no long-run effect of reduced class size. However, our study confirms that the long-run effect of class size seems to be qualitatively similar to the short-run effect on student achievement. While there appears to be positive effects of smaller class size both in term of student achievement, educational attainment, and income in contexts analyzed in the United States, Sweden and Denmark, there appears to be no effect on student achievement, educational attainment or income within the institutional setting of Norway.

The difference between our results and the other Scandinavian countries is of special interest since these countries are viewed as very similar. One potential explanation for the different results is that school districts are generally much smaller in Norway than in Sweden and Denmark.[29] However, our finding that the class size effect in Norway does not depend on school district size speaks against this explanation. Another possibility might be that teacher quality differs systematically between countries.[30] The absence of robust significant interaction effects between class size and our indicator for teacher quality does not support this explanation either.

If schools use compensatory policies and increase the use of other inputs in grades with larger classes, such as teacher assistants, the estimated class size effect will be biased towards zero. Unfortunately, other measures on education inputs than the class size are not available for the time period of the present paper. The potential for such policies

[29]Both in Sweden, Denmark and Norway, the municipalities (school districts) are multi-purpose local governments with the major responsibility for local welfare services. A major consolidation reform in Sweden in 1974 reduced the number of municipalities to about 280, while Denmark in 2007 implemented a consolidation of municipalities from 271 to 98. In contrast, Norway has about 440 municipalities even though the population in 1990 (4.2 mill) was half of that in Sweden and roughly 20% lower than in Denmark. Average municipality size in 1990 was around 30,000, 19,000 and 10,000 in Sweden, Denmark and Norway, respectively.

[30]The share of teachers certified for teacher jobs varies substantially between Norway and Sweden. According to Andersson, Johansson, and Waldenström (2011), more than 15 percent of the Swedish teachers were non-certified on average in 2000, while Bonesrønning, Falch, and Strøm (2005) show that the corresponding number for Norway is about 6 percent.

used to be low, but has increased over time by increased school budgets, availability of computers, and due to school budgets being uncoupled with the class size rule to an increasing degree. Using data on the number of teachers for the school years 2002-2003, Leuven, Oosterbeek, and Rønning (2008) investigate whether input substitution can explain the absence of any class size effect in the short run. They find only week evidence of input substitution, and that such a substitution cannot drive the results. In addition, it is unlikely that any compensatory policies towards large classes should be different across the Scandinavian countries. At a general level, the class size effects might obviously also depend on characteristics of the students, although such characteristics vary to a smaller degree across countries. First, there is some evidence that the class size effect is largest at young ages. Ehrenberg, Brewer, Gamoran, and Willms (2001) hypothesize that small classes during the elementary grades develop working habits that enable students to take advantage of learning opportunities in later grades. The STAR experiment was targeted towards students up to third grade. Fredriksson, Öckert, and Oosterbeek (2013) investigate class size effects at ages 11-13. However, several papers find a positive effect of resources also in higher grades. Fredriksson and Öckert (2008) find for Sweden a positive effect of the teacher/student ratio on student performance at age 16 in a difference-in-differences framework. For Denmark, Browning and Heinesen (2007) find that lower class size in grade 8 increases the probability of completing high school and years of education, and Heinesen (2010) finds a positive effect of subject-specific class size in lower secondary education in a student fixed effects framework. In addition, Leuven and Løkken (2015) find no long-run effect of class size in primary education in Norway. This evidence clearly suggests that our use of class size in lower secondary education (grades 8-10) cannot explain the different results between Norway and the other Scandinavian countries.

A final issue is that class size effects may differ across students with different socioeconomic characteristics. First, there is evidence of gender differences in competitiveness (Buser, Niederle, Oosterbeek, et al., 2014), which might give gender differences in the class size effects. Larger classes arguably have a more competitive environment. However, also for gender differences, the evidence is mixed for class size reductions. In separate analyses reported in Appendix Table A3, we do not find different class size effects for males and females in the Norwegian data.

Second, small classes might be most beneficial for students with disadvantaged backgrounds, who do not have the same resources in the home to support their education as other students. This is the typical finding from the STAR experiment (Dynarski, Hyman, and Schanzenbach, 2013) and other studies (Bosworth, 2014; Vaag Iversen and Bonesrønning, 2013). On the other hand, Fredriksson, Öckert, and Oosterbeek (2013) find strongest class size effects for students with high parental income. Appendix Table A3 shows that we do not find evidence of such heterogeneity, which is consistent with the findings for variation in student outcomes in Table 4 above.

8 Conclusion

The lack of conclusive evidence on the effect of school resources on student test scores calls for systematic studies of possible heterogeneous effects using credible identification strategies. This paper uses rich register data from Norway from a long time period combined with a quasi-experimental empirical strategy to estimate both the average effect of class size and to which extent the effect varies with a range of external conditions facing schools and students. Using a strict class size rule in an RDD framework, we first show that on average there is no evidence that lower class size increases long-run outcomes as earnings and educational attainment. This is in accordance with the previous Norwegian results for short-run outcomes.

Second, we investigate heterogeneity in class size effects by interacting class size with indicators of teacher quality, the extent of upper secondary school choice, school district size, local fiscal constraints, and labor market conditions within the same quasi-experimental framework. Overall, we find that class size effects do not depend on such external conditions.

Our results stand in sharp contrast to experimental evidence from the United States and quasi-experimental evidence from Sweden and Denmark finding significant and numerically important positive effects of reduced class size on both short-run and long-run outcomes. The absence of interaction effects with measured external conditions indicate that between country differences in teaching practices and educational culture are relevant explanations for the different results.

References

Adkins, L. C., and R. L. Moomaw (2003): "The impact of local funding on the technical efficiency of Oklahoma schools," *Economics Letters*, 81(1), 31–37.

Altinok, N., and G. Kingdon (2012): "New Evidence on Class Size Effects: A Pupil Fixed Effects Approach*," *Oxford Bulletin of Economics and Statistics*, 74(2), 203–234.

Andersson, C., P. Johansson, and N. Waldenström (2011): "Do you want your child to have a certified teacher?," *Economics of Education Review*, 30(1), 65–78.

Angrist, J., E. Battistin, and D. Vuri (2015): "In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno," IZA Discussion Papers 8959, Institute for the Study of Labor (IZA).

Angrist, J. D., and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114(2), 533–575.

Anzia, S. F. (2011): "Election timing and the electoral influence of interest groups," *The Journal of Politics*, 73(02), 412–427.

Barankay, I., and B. Lockwood (2007): "Decentralization and the productive efficiency of government: Evidence from Swiss cantons," *Journal of Public Economics*, 91(5), 1197–1218.

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

Berry, C. R., and M. R. West (2010): "Growing pains: The school consolidation movement and student outcomes," *Journal of Law, Economics, and Organization*, 26(1), 1–29.

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

Björklund, A., P.-A. Edin, P. Frerikssson, and A. B. Krueger (2004): "Education, equality and efficiency: An analysis of Swedish school reforms during the 1990s," *IFAU report*, 1, 108–109.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2013): "Under pressure? The effect of peers on outcomes of young adults," *Journal of Labor Economics*, 31(1), 119–153.

Bonesrønning, H. (2003): "Class size effects on student achievement in Norway: Patterns and explanations," *Southern Economic Journal*, pp. 952–965.

——— (2013): "Public employees and public sector reform implementation," *Public Choice*, 156(1-2), 309–327.

Bonesrønning, H., T. Falch, and B. Strøm (2005): "Teacher sorting, teacher quality, and student composition," *European Economic Review*, 49(2), 457–483.

Borge, L.-E., and J. Rattsø (2008): "Property taxation as incentive for cost control: Empirical evidence for utility services in Norway," *European Economic Review*, 52(6), 1035–1054.

Bosworth, R. (2014): "Class size, class composition, and the distribution of student achievement," *Education Economics*, 22(2), 141–165.

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

Buser, T., M. Niederle, H. Oosterbeek, et al. (2014): "Gender, competitiveness and career choices," *Quarterly Journal of Economics*, 129(3), 1409–1447.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics*, 126(4), 1593–1660.

Chubb, J. E., and T. M. Moe (1988): "Politics, markets, and the organization of schools," *American Political Science Review*, 82(04), 1065–1087.

Clark, D. (2011): "Do Recessions Keep Students in School? The Impact of Youth Unemployment on Enrolment in Post-compulsory Education in England," *Economica*, 78(311), 523–545.

Coleman, J. S., E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld, and R. York (1966): *Equality of Educational Opportunity*.

Costrell, R. M. (1994): "A simple model of educational standards," *American Economic Review*, pp. 956–971.

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

Dixit, A. (2002): "Incentives and organizations in the public sector: An interpretative review," *Journal of Human Resources*, pp. 696–727.

Dixit, A. K. (1998): *The making of economic policy: a transaction-cost politics perspective*. MIT press.

Dobbelsteen, S., J. Levin, and H. Oosterbeek (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.

Driscoll, D., D. Halcoussis, and S. Svorny (2003): "School district size and student performance," *economics of Education Review*, 22(2), 193–201.

Dynarski, S., J. Hyman, and D. W. Schanzenbach (2013): "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion," *Journal of Policy Analysis and Management*, 32(4), 692–717.

Ehrenberg, R. G., D. J. Brewer, A. Gamoran, and J. D. Willms (2001): "Class size and student achievement," *Psychological Science in the Public Interest*, pp. 1–30.

Falch, T., and J. A. Fischer (2012): "Public sector decentralization and school performance: International evidence," *Economics Letters*, 114(3), 276–279.

Falch, T., K. Johansen, and B. Strøm (2009): "Teacher shortages and the business cycle," *Labour Economics*, 16(6), 648–658.

Falch, T., and L. R. Naper (2013): "Educational evaluation schemes and gender gaps in student achievement," *Economics of Education Review*, 36, 12–25.

Figlio, D., and C. Hart (2014): "Competitive effects of means-tested school vouchers," *American Economic Journal: Applied Economics*, 6(1), 133–156.

Fiva, J. H., and M. Rønning (2008): "The incentive effects of property taxation: Evidence from Norwegian school districts," *Regional Science and Urban Economics*, 38(1), 49–62.

Fredriksson, P., and B. Öckert (2008): "Resources and Student Achievement - Evidence from a Swedish Policy Reform," *Scandinavian Journal of Economics*, 110(2), 277–296.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013): "Long-Term Effects of Class Size*," *Quarterly Journal of Economics*, 128(1), 249–285.

Gelman, A., and G. Imbens (2014): "Why high-order polynomials should not be used in regression discontinuity designs," Discussion paper, National Bureau of Economic Research.

Glaeser, E. L. (1996): "The incentive effects of property taxes on local governments," *Public Choice*, 89(1), 93–111.

Gunderson, M. (2005): "Two faces of union voice in the public sector," *Journal of Labor Research*, 26(3), 393–413.

Hægeland, T., O. Raaum, and K. G. Salvanes (2012): "Pennies from heaven? Using exogenous tax variation to identify effects of school resources on pupil achievement," *Economics of Education Review*, 31(5), 601–614.

Hanushek, E. A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24(3), 1141–77.

——— (2003): "The Failure of Input-based Schooling Policies*," *Economic Journal*, 113(485), F64–F98.

——— (2006): "School resources," *Handbook of the Economics of Education*, 2, 865–908.

Haraldsvik, M. (2003): "Inntaksprosedyrer for den videregående skole: Grad av valgfrihet," Institutt for samfunnsøkonomi, NTNU.

——— (2012): "Does performance based school choice affect student achievement?," Discussion paper, Doctoral thesis at NTNU 2012:346.

Hattie, J. (2005): "The paradox of reducing class size and improving learning outcomes," *International Journal of Educational Research*, 43(6), 387–425.

Heinesen, E. (2005): "School district size and student educational attainment: evidence from Denmark," *Economics of Education Review*, 24(6), 677–689.

——— (2010): "Estimating Class-size Effects using Within-school Variation in Subject-specific Classes*," *Economic Journal*, 120(545), 737–760.

Hoxby, C. M. (1996): "How teachers' unions affect education production," *Quarterly Journal of Economics*, pp. 671–718.

Hoxby, C. M. (1999): "The productivity of schools and other local public goods producers," *Journal of Public Economics*, 74(1), 1–30.

——— (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, pp. 1239–1285.

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

Krueger, A. B. (2003): "Economic considerations and class size*," *Economic Journal*, 113(485), F34–F63.

Krueger, A. B., and D. M. Whitmore (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.

Lavecchia, A. M., H. Liu, and P. Oreopoulos (2014): "Behavioral economics of education: Progress and possibilities," Discussion paper, National Bureau of Economic Research.

Lazear, E. P. (2001): "Educational Production," *Quarterly Journal of Economics*, 116(3), 777–803.

Lee, D. N. (2013): "The impact of repealing Sunday closing laws on educational attainment," *Journal of Human Resources*, 48(2), 286–310.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

Leuven, E. (2013): "Long term impacts of class size in middle school and integrated primary schools," Mimeo.

Leuven, E., and S. A. Løkken (2015): "Long term impacts of class size in compulsory schooling.," Mimeo.

Leuven, E., H. Oosterbeek, and M. Rønning (2008): "Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway*," *Scandinavian Journal of Economics*, 110(4), 663–693.

Lindahl, M. (2005): "Home versus school learning: A new approach to estimating the effect of class size on achievement," *Scandinavian Journal of Economics*, 107(2), 375–394.

Loeb, S., and K. Strunk (2007): "Accountability and Local Control: Response to Incentives with and without Authority over Resource Generation and Allocation," *Education Finance and Policy*, 2(1), 10–39.

Lott, J., and L. W. Kenny (2013): "State teacher union strength and student achievement," *Economics of Education Review*, 35, 93–103.

Lovenheim, M. F. (2009): "The effect of teacher' unions on education production: Evidence from union election certifications in three midwestern states," *Journal of Labor Economics*, 27(4), 525–587.

Machin, S., and K. G. Salvanes (2016): "Valuing School Quality via a School Choice Reform," *Scandinavian Journal of Economics*, 118(1), 3–24.

Mensah, Y. M., M. P. Schoderbek, and S. P. Sahay (2013): "The effect of administrative pay and local property taxes on student achievement scores: Evidence from New Jersey public schools," *Economics of Education Review*, 34, 1–16.

Moe, T. M. (2001): "A union by any other name," *Education Next*, 1(3).

——— (2011): *Special interest: Teachers unions and America's public schools*. Brookings Institution Press.

Mueller, S. (2013): "Teacher experience and the class size effect - Experimental evidence," *Journal of Public Economics*, 98, 44–52.

OECD (2011): "OECD Reviews of evaluation and assessment in education: Norway.," http://www.oecd.org/norway/48632032.pdf.

Rattsø, J., and R. J. Sørensen (2004): "Public employees as swing voters: Empirical evidence on opposition to public reform," *Public Choice*, 119(3), 281–310.

Reiling, R. B., and B. Strøm (2015): "Upper secondary school completion and the business cycle," *Scandinavian Journal of Economics*, 117(1), 195–219.

Rockoff, J. (2009): "Field experiments in class size from the early twentieth century," *Journal of Economic Perspectives*, pp. 211–230.

Strunk, K. O., and J. A. Grissom (2010): "Do strong unions shape district policies? Collective bargaining, teacher contract restrictiveness, and the political power of teachers' unions," *Educational Evaluation and Policy Analysis*, 32(3), 389–406.

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

Vaag Iversen, J. M., and H. Bonesrønning (2013): "Disadvantaged students in the early grades: will smaller classes help them?," *Education Economics*, 21(4), 305–324.

Webbink, H. D. (2005): "Causal effects in education," *Journal of Economic Surveys*, 19(4), 535–560.

Wössmann, L., and M. West (2006): "Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS," *European Economic Review*, 50(3), 695–736.

A Appendix

Table A1: Data reduction

	Observations	Reduction	% Reduction
1. Sample 1982-2000 (without 1990)	1,040,840		
2. Non-missing class size	1,003,149	37,691	3,62 %
3. 16 years old when graduating from lower secondary school	953,512	49,637	4,95 %
4. At least 10 school observations	953,183	329	0,03 %
5. Non missing years of education	952,514	669	0,07 %
5. Non missing log of income	903,828	49,355	5,18 %

Note: Data on the school identifier is missing in 1990. 49,355 observations have zero income, which are excluded from the analysis because we use the logarithmic value of income.

Table A2: Main results with socioeconomic characteristics and enrollment controls

	Log income		Years of education	
	(1)	(2)	(3)	(4)
Average class size grades 8-10	-0.000396	-2.35e-05	0.000677	-0.00117
	(0.0007)	(0.0021)	(0.0021)	(0.0069)
Girl	-0.336***	-0.337***	0.545***	0.526***
	(0.0033)	(0.0052)	(0.0081)	(0.0148)
Parental education: High School	0.113***	0.119***	1.049***	1.092***
	(0.0025)	(0.0054)	(0.0098)	(0.0182)
Parental education: Bachelor	0.163***	0.161***	2.391***	2.495***
	(0.0033)	(0.0070)	(0.0121)	(0.0226)
Parental education: Masters +	0.176***	0.181***	3.349***	3.485***
	(0.0051)	(0.0107)	(0.0158)	(0.0272)
Parental education: Unknown	0.00863	0.0266**	0.750***	0.831***
	(0.0062)	(0.0133)	(0.0250)	(0.0581)
First generation immigrant	-0.0575***	-0.0642***	-0.0318	-0.184***
	(0.0104)	(0.0196)	(0.0365)	(0.0673)
Second generation immigrant	0.0305**	0.00807	0.454***	0.164
	(0.0146)	(0.0318)	(0.0438)	(0.1008)
Only mother working	0.0423***	0.0349***	0.129***	0.107***
	(0.0026)	(0.0056)	(0.0092)	(0.0188)
Only father working	0.0393***	0.0345***	0.0535***	0.0588***
	(0.0026)	(0.0055)	(0.0094)	(0.0189)
Both parents working	0.108***	0.112***	0.458***	0.476***
	(0.0024)	(0.0050)	(0.0091)	(0.0172)
Birth month	0.000681***	0.000235	0.00602***	0.00643***
	(0.0002)	(0.0005)	(0.0007)	(0.0017)
Enrollment	-0.00103	0.00242	0.00942***	-0.0103
	(0.0010)	(0.0025)	(0.0033)	(0.0087)
Enrollment2	1.59e-05		-0.000117**	
	(0.0000)		(0.0001)	
Enrollment3	-9.78e-08		6.51e-07*	
	(0.0000)		(0.0000)	
Enrollment4	2.09e-10		-1.25e-09*	
	(0.0000)		(0.0000)	
Segment 1	-0.000288	0.600	0.0166	-2.441
	(0.0094)	(0.6228)	(0.0307)	(2.1299)
Segment 2	0.00225	0.530	-0.0181	-2.142
	(0.0127)	(0.5421)	(0.0409)	(1.8566)
Segment 3	0.00136	0.458	0.00627	-1.802
	(0.0151)	(0.4635)	(0.0472)	(1.5851)
Segment 4	0.00230	0.384	0.0105	-1.463
	(0.0172)	(0.3852)	(0.0532)	(1.3167)
Segment 5	0.0162	0.304	-0.0220	-1.197
	(0.0195)	(0.3079)	(0.0607)	(1.0556)
Segment 6	0.0407*	0.243	-0.0478	-0.934
	(0.0241)	(0.2312)	(0.0747)	(0.7955)
Segment 7	0.0383	0.199	-0.113	-0.505
	(0.0307)	(0.1515)	(0.0985)	(0.5504)
Segment 8	0.00253	0.0722	-0.0265	-0.0854
	(0.0467)	(0.0771)	(0.1530)	(0.2659)
Segment 9	-0.0374		0.323	
	(0.0861)		(0.2947)	
R-squared	0.107	0.110	0.151	0.172
Predicted class size (the instrument), first stage	0.56***	0.41***	0.56***	0.40***
	(0.013)	(0.023)	(0.013)	(0.023)
F-value first stage	1,935	295.9	1,934	294.7
R-squared first stage	0.4906	0.4864	0.4893	0.4855
Observations	903,828	170,604	952,514	179,799
No. of schools	1,156		1,156	
Enrollment controls	Pol. and seg. FE	Linear and seg. FE	Pol. and seg. FE	Linear and seg. FE
Subsample +/- 3 students	No	Yes	No	Yes
School FE	Yes	No	Yes	No

Note: All regressions include socioeconomic characteristics and time/age fixed effects. Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001. Standard errors are clustered at the school level. Socioeconomic characteristics are described in section 3.1. When describing enrollment controls, seg. is segment and pol. is polynomial.

Table A3: Subsample analysis

	Log income		Years of education	
	(1)	(2)	(3)	(4)
A. Girls				
Average class size grades 8-10	0.000188	0.00227	-0.000529	0.00411
	(0.0009)	(0.0027)	(0.0029)	(0.0089)
Observations	443,057	83,593	466,957	88,101
B. Boys				
Average class size grades 8-10	-0.000916	-0.00186	0.00279	-0.00481
	(0.0009)	(0.0027)	(0.0029)	(0.0091)
Observations	460,770	87,011	485,557	91,698
C. Parental education more than high school				
Average class size grades 8-10	-0.000730	-0.000319	-0.000983	-0.00162
	(0.0010)	(0.0029)	(0.0013)	(0.0042)
Observations	392,413	74,591	405,286	77,021
D. Parental education less than high school				
Average class size grades 8-10	-0.000376	0.000225	-0.00170	0.00330
	(0.0008)	(0.0024)	(0.0016)	(0.0048)
Observations	511,412	96,013	547,225	102,778
E. Immigrant				
Average class size grades 8-10	-0.00764	0.00465	-0.000724	-0.0357
	(0.0073)	(0.0165)	(0.0195)	(0.0443)
Observations	15,206	2,977	17,427	3,364
F. Non immigrant				
Average class size grades 8-10	-0.000421	-0.000142	0.00116	-0.000582
	(0.0007)	(0.0021)	(0.0021)	(0.0069)
Observations	888,502	167,627	934,968	176,435
Enrollment controls	Pol. and seg. FE	Linear and seg. FE	Pol. and seg. FE	Linear and seg. FE
School FE	Yes	No	Yes	No
Subsample 3+/-	No	Yes	No	Yes

Note: For enrollment controls, seg. is segment and pol. is polynomial.

Chapter 3:

Municipality mergers

Astrid Marie Jorde Sandsør and Bjarne Strøm

Municipality mergers*

Astrid Marie Jorde Sandsør† Bjarne Strøm ‡

Abstract

To merge municipalities is an important policy issue in many countries, yet empirical evidence on the effect of municipality size on the production and quality of local public services is scarce. We use the spatial and temporal variation in forced municipality merges in a difference-in-differences approach to provide quasi-experimental evidence of the effect of municipality size on school output, measured by student educational attainment and income in adulthood. We find that municipality mergers increase student income by 2-3%, while the effect on educational attainment is less clear.

Keywords: school district, quasi-experiment, educational attainment, income
JEL codes: I2, H7

*Thanks to Torberg Falch and Kalle Moene as well seminar participants at the Annual Meeting of the Norwegian Association of Economists and the Department of Economics, University of Oslo for helpful comments and suggestions

†Department of Economics, University of Oslo
‡Department of Economics, Norwegian University of Science and Technology

1 Introduction

The size and number of local governments is an important policy question. Municipal amalgamation reforms and consolidation of school districts are hot issues in many countries and such reforms are currently on the political agenda in countries like Norway and Finland.[1] While fiscal decentralization is generally believed to be beneficial for society as suggested by the decentralization theorem formulated by Oates (1972), common arguments for amalgamation reforms are based on economics of scale, that increased school district size implies reduced expenditure per pupil. However, the size effect on output quality is not obvious. Expenditure reduction may come at the cost of reduced quality of services provided by the local units. On the one hand, larger local units may decrease local autonomy at the provider level (school, day care institution or homes for elderly). If the population becomes more heterogeneous as a result, the larger local governments might be less able to meet the needs of the heterogeneous users of public services. On the other hand, it is possible that larger local governments will have more professional administration and management of resources and so increase output quality for a given amount of resources available. For example, the probability of hiring professional and able school administrators may be higher in large than in small school districts. Ultimately, the relationship between local government size and output quality can only be resolved by empirical studies.

Below we investigate the effect of municipal size on educational output in terms of student educational attainment and earnings in adulthood using rich data from administrative registers in Norway. To provide credible evidence, we explore the spatial and temporal variation in municipal size from enforced municipality mergers taking place in Norway in the 1980's and 1990's in a difference-in-differences approach. Using outcomes in terms of educational attainment and earnings has several advantages when studying the relationship between municipality size and output quality. First, educational services in terms of compulsory schooling is provided by all municipalities, small and large. The users are well defined (children age 7-16) and to the extent that private schooling is not an option, services are solely provided by the local public sector. Second, educational attainment and earnings in adulthood may be more relevant measures of education output than test scores often used in estimates of education production functions as these broader measures are more likely to reflect the multi-dimensional property of educational production.

[1] Municipal merger reforms have been implemented in a number of countries including Canada (Dafflon, 2013), Denmark (Hansen, 2014), Sweden (Hinnerich, 2009; Jordahl and Liang, 2010), Israel (Reingewertz, 2012) and to some extent in Finland (Saarimaa and Tukiainen, 2015).

Third, we can control for individual socioeconomic characteristics in the analysis. Lastly, we are able to use a school fixed effects strategy. To the extent that municipality mergers did not lead to school consolidation, we can compare students before and after the merger attending the same schools.

Causal evidence of the output-size relationship is hard to obtain for a number of reasons. The size of a local unit measured by the number of inhabitants as an explanatory variable in traditional expenditure or output equations is clearly endogenous since fiscal variables and the production and quality of local public services affect migration decisions. An obvious alternative is to explore municipality or school district mergers in a quasi-experimental framework. However, to the extent that mergers are voluntary, endogeneity issues are still a concern. For municipalities to merge voluntarily, they not only need to find that the benefits outweigh the costs, they also must overcome any political coordination problems. Central authorities might have more knowledge about the expected benefits of a merger and can overcome coordination problems by enforcing the merger, making these mergers especially interesting to study. Using large structural reforms induced by the central government as the reform in Sweden in the 1950's or the reform in Norway in the 1960's can potentially offer better identifying opportunities. However, such large structural reforms often occur in combination with other reforms in the provision of local public services making it difficult to disentangle the impact of the different reform elements.[2] This paper uses forced mergers from a period without other large national structural reforms in the provision of local services and therefore offers a better opportunity to isolate the effect of mergers on municipal output.

The mergers we study were enforced by the central authorities based on recommendations from two official Norwegian reports (Norwegian Ministry of Local Government and Labor, 1986, 1989).[3] The mergers were former city municipalities merging with surrounding municipalities, having two main benefits. First, it creates a natural comparison group of city and surrounding municipalities. Second, there is reason to believe that merging could have different consequences for the city and surrounding municipalities. The mergers were often met by large local resistance in the municipalities surrounding the city and several referenda gave very little support

[2]For example the large reduction in the number of municipalities in Norway in the 1960's coincided with substantial changes in the education system (extension of mandatory school years from 7 to 9, a new curriculum and a new tracking system in the new compulsory lower secondary school, see Aakvik, Salvanes, and Vaage (2010). Similarly, the 1952 reform in Sweden which drastically reduced the number of municipalities coincided with extension of mandatory school years from 7 to 9, see Meghir and Palme (2005).

[3]All recommended mergers were carried out except in the case of the city municipality Hamar, where the merge met such large resistance from Løten municipality that they managed to remain independent.

for merger plans. If this resistance reflected correct anticipations of future merger effects on service production, the effect on output and quality in schools located in former surrounding municipalities could be negative. The rich individual by school by municipality data available to us, makes it possible to test this hypothesis.

Partly because of the large local resistance in the merger process, central authorities decided to no longer enforce mergers after the last merger was carried out in 1994. Although the municipalities chosen to merge are not random, the timing of the mergers might be. Also, there might have been municipalities that were next in line when the central authorities decided to abandon enforced mergers. This creates some randomness to the selection and timing and further strengthens our analysis.

This paper estimates the effect of school district size through municipal mergers using a difference-in-differences approach with a school fixed effects strategy. Municipality mergers are found to significantly increase student income in adulthood by 2-3%, while the effect on educational attainment is generally positive, but not precisely estimated. To enhance the understanding of possible mechanisms behind this important result, we further investigate possible heterogeneous effects by school location and the effect of mergers on school characteristics and fiscal variables, using the same difference-in-differences approach but with municipalities as the unit of analysis.

Our results clearly show that the income effect is driven by students enrolled in schools in pre-merger municipalities surrounding the former city. The effect on students enrolled in schools located in the pre-merger city is numerically very small and far from significant. Thus, the hypothesis that former surrounding municipalities resisted merger because of correct anticipations of negative future merger effects on service production and quality is not supported by the empirical results. Rather the evidence suggests the opposite. Output and quality as measured by our variables increased in these former surrounding municipalities. The former cities became administrative centers in the new municipalities. The finding is consistent with the hypothesis that students enrolled in schools in former surrounding municipalities took advantage of potential gains in existing administrative quality in the former cities, although further research is needed to confirm this interpretation.

We also find that the merger reduced total municipal expenditure per capita by nearly 5% which is qualitatively consistent with the evidence in Reingewertz (2012) although numerically smaller. The effect on expenditure per student (6-15 years old) is also negative but not statistically significant. This suggests that the positive student income effect in adulthood cannot be explained by increased total budgets in merged municipalities or budget reallocation in favor of the education sector.

Finally, we find that the number of lower secondary schools, the number of persons aged 7-16 and overall teacher quality measured by the share of teachers without a teacher certification at the municipality level is not significantly affected by the merger. Thus, we tentatively conclude that systematic changes in the number of schools, cohort size and teacher quality cannot explain the income effect.

The paper is organized as follows. Section 2 presents a review of the literature on the optimal size of local public authorities and relevant empirical studies. Section 3 describes the institutions and data while the identification and model specification are presented in Section 4. Section 5 presents the main results of the difference-in-differences estimation of municipality mergers on log income and years of education. Section 6 presents various robustness checks and Section 7 presents a discussion of mechanisms. Section 8 concludes.

2 Theoretical background and empirical literature

2.1 Theoretical background

The first generation fiscal federalism literature, represented by Oates' seminal contribution (Oates, 1972), formulated what is called the decentralization theorem. This theorem states that public services which are local in nature should be produced and financed at the local level because these entities can meet the demands of the local population in the least costly way.[4] Moreover, from a different perspective, Tiebout (1956) showed that an optimal allocation of private and public goods can be reached when households sort themselves across jurisdictions according to their preferences for local services and local taxes. Endogenous formation of a large number of jurisdictions and household mobility are central mechanisms to reach the Tiebout equilibrium.

The early theoretical contributions have been extended and challenged by authors taking political issues into account. On the one hand, authors in the public choice tradition, represented by the seminal contribution by Brennan and Buchanan (1980), also view fiscal decentralization as beneficial, but for a very different reason. In their view, the public sector acts as an agent ("Leviathan") with the objective of maximizing revenues extracted from the private sector. In this perspective decentralization of taxing and production decisions creates competition between local jurisdictions and leads to enhanced economic efficiency and taming of the "Leviathan". In both the Tiebout and the public choice model, enforced mergers of local jurisdic-

[4]This view is also presented in Musgrave and Musgrave (1973) and Atkinson and Stiglitz (1980).

tions could lead to a less efficient production of local services.

The second generation fiscal federalism literature has extended the original approach in Oates (1972) with an explicit modelling of the political process both at the central and local government level (see Oates (2005) for an extensive review). While the first generation literature assumes that central provision requires a uniform level of public output, recent authors allow for varying levels of outputs across jurisdictions in a centralized regime. For example, Lockwood (2002) and Besley and Coate (2003) model the centralized outcome as a vector of local outcomes determined by locally elected representatives. In their framework, decentralization has additional benefits in terms of reduced corruption, waste and poor governance compared to a centralized regime. These benefits must be weighed against potential losses due to spillovers between jurisdictions and scale effects in the production of local services.[5] Alesina and Spolaore (1997) explicitly consider jurisdictions with heterogeneous populations and argue that there is a trade-off between the benefits of large political jurisdictions and the costs of heterogeneity in large populations. They find that the democratic process leads to an inefficiently large number of jurisdictions (countries). Alesina, Baqir, and Hoxby (2004) take a similar approach and provide empirical evidence from U.S. municipalities, school districts and special districts that a trade-off between size and heterogeneity exists. They find a negative relationship between local government size and racial and income heterogeneity while no relationship is found between size and religious or ethnic heterogeneity.

2.2 Empirical literature

The theoretical models discussed above, suggest that gains from decentralization of public service production to a large number of jurisdictions must be balanced against potential economies of scale. While some studies confirm the existence of economies of scale in most municipal services,[6] other studies find that they only exist up to a certain size,[7] or find no correlation between costs and size.[8] However, local authorities have many services and optimal size may differ according to service. Most of the existing empirical literature has concentrated on scale effects on fiscal outcomes, such as expenditures and taxes. Oates (1985) provides an empirical test of the hypotheses that more decentralization reduces the size of government and

[5]Other papers in this literature are Besley and Case (1995), Ellingsen (1998) and Coate and Knight (2007).

[6]Kraus (1981); Duncombe and Yinger (2007); Razin (1999); Callan and Thomas (2001); DeBoer (1992); Farsi, Fetz, and Filippini (2007)

[7]Reiter and Weichenrieder (1997); Solé-Ollé and Bosch (2005); Breunig and Rocaboy (2008)

[8]Gyimah-Brempong (1987); Derksen (1988)

the tax burden as predicted by the public choice view represented by Brennan and Buchanan (1980). He finds no clear evidence that countries with more decentralized government structure have lower total public expenditure. Zax (1989) using data from U.S. local governments finds mixed evidence. While the size of multipurpose local governments like municipalities is negatively associated with measures of fiscal decentralization, the opposite seems to be the case for single-purpose governments like school districts. While potential effects on fiscal variables are interesting, knowledge of the relationship between local public output and quality, and size of political jurisdictions is warranted, but few empirical studies exist on this relationship.

One recent study, building explicitly on the fiscal federalism literature and providing evidence on the effect of decentralization on public output, is Barankay and Lockwood (2007). Using panel data for Swiss cantons, they find that educational attainment is higher in cantons with more decentralized provision of educational services measured by the share of education expenditures in a canton provided at the county level.

A small literature has also studied the effects of school district size on school output in a traditional educational production framework. The evidence on the effect of district size on student performance in this literature is mixed. Driscoll, Halcoussis, and Svorny (2003) use data from California to estimate an educational production function with test scores as output and find a negative effect of district size on test scores. Andrews, Duncombe, and Yinger (2002) review five studies from the United States that estimate the returns to school district size using test scores as the dependent variable. Of these, Walberg and Fowler (1987) and Ferguson (1991) find a negative effect of district size on test scores, Sebold and Dato (1981) and Baum (1986) find no or positive effects of district size, while Ferguson and Ladd (1996) find positive effects of district size. Kiesling (1967), Niskanen (1998) and Jacques, Brorsen, and Richter (2000) all find negative effects of district size on test scores.[9]

Test scores could be misleading as a measure of quality of school outputs, as they are possible to manipulate (Angrist, Battistin, and Vuri, 2015) and only measure cognitive skills, while non-cognitive skills might also be important for future outcomes (Kautz, Heckman, Diris, ter Weel, and Borghans, 2014). Both arguments suggest that analyses of long-run outcomes in terms of educational attainment and income provide the most credible evidence of the effect of district size (Driscoll, Halcoussis, and Svorny, 2003). Heinesen (2005) analyzes the effect of size of school district on educational attainment using Danish administrative register data and

[9]See also Fox (1981)

finds that educational attainment is higher for students from larger districts, i.e. districts with population above 15,000.

A problem with the studies above is that smaller and larger districts differ in characteristics that are not well measured. Over time, highly effective schools and districts may attract more students which will generate a bias towards finding increasing returns to size. Berry and West (2010) attempt to address this concern by exploiting the variation in the timing of consolidation across the United States to estimate the effect of changing school and district size on student outcomes. They find that larger districts have some modest gains with respect to returns to education but that these gains are outweighed by the harmful effect of larger schools. Reingewertz 2012 uses a difference-in-differences methodology to study the Israeli municipality consolidation reform of 2003 and finds positive effects of consolidations, among other things on the share of matriculation exam recipients. Gordon and Knight (2008) use school district consolidations to examine the effect of whole-grade sharing and consolidation of school districts on pupil-teacher ratio, enrollment, drop-out, revenues, and local expenditures, and their findings suggests an absence of efficiency gains from consolidations.

Other studies have looked at the effect of school consolidation on student outcomes. While not directly related to school district or municipality size, school consolidation may be one channel whereby municipality mergers can affect student outcomes. Beuchert, Humlum, Nielsen, and Smith (2015) exploit exogenous variation in school consolidations in Denmark to analyze their impact on student achievement and find that school consolidations have negative effects in the short run that are more pronounced for the students experiencing a school closure. Berry and West (2010) find that students educated in states with small schools have higher returns to education and complete more years of schooling.[10]

The methodology in this paper is similar to that of many other papers studying the impact of municipality mergers on various outcomes. Saarimaa and Tukiainen (2015) use a difference-in-differences methodology to investigate the free riding behavior in relation to voluntary municipal mergers and find that stronger free riding incentives create increased debt and spending. Reingewertz (2012) uses a difference-in-differences methodology to study the Israeli municipality consolidation reform of 2003 and finds that municipality consolidation reduced municipal expenditures without lowering the level of services. Moisio and Uusitalo (2013) investigates the impact

[10]See also Kuziemko (2006); Schwartz, Stiefel, and Wiswall (2013); Abdulkadiroğlu, Hu, and Pathak (2013); de Haan, Leuven, and Oosterbeek (2014); Humlum and Smith (2015); Barrow, Schanzenbach, and Claessens (2015); Engberg, Gill, Zamarro, and Zimmer (2012); Brummet (2014); Liu, Zhang, Luo, Rozelle, and Loyalka (2010).

of municipal mergers on local public expenditures in Finland. Rather than use a difference-in differences methodology, they use matching to compare pairs of merged municipalities to similar pairs of unmerged municipalities. The municipalities mergers they study are voluntary municipalities, and this method attempts to control for the non-random selection of municipalities that chose to merge.

3 Institutions and data

3.1 School system

Compulsory education is one of the core responsibilities of the Norwegian municipalities. The relative importance of the education sector in municipality activity is illustrated by its budget share of 43% on average for the 1980-1990 period, while the corresponding shares for child care, health care, culture and infrastructure is 4%, 18%, 6% and 17% respectively, see Borge, Brueckner, and Rattsø (2014). Schooling is provided free of charge and only a very small fraction of children enroll in private schools. Compulsory education in Norway consists of primary school and lower secondary school, and ends the year students turn 16 years of age.[11] Most students continue on to upper secondary education, which is divided into a three-year long academic study track and different vocational study tracks. After a major reform in 1994, vocational study tracks typically last for four years (including two years of apprenticeship training). Acceptance to upper secondary school is based on the grades achieved in grade 10. However, all students have been guaranteed admission to upper secondary education since 1994.

There is no possibility to fail a class in primary or in lower secondary education during the empirical period, which implies that all students finish compulsory education on time.[12] Education is comprehensive with a common curriculum for all students and there is no tracking. The cutoff between grades is birth at January 1.

[11] During the empirical period, the school starting age was 7 years. In 1997 the school starting age was reduced from 7 to 6 years such that today primary education consists of grades 1-7 (ages 6-13) and lower secondary education consists of grades 8-10 (ages 14-16). We refer to grades 8-10 as lower secondary education throughout the paper.

[12] In some cases, students do not start primary education at the expected age, which implies that they finish lower secondary education at a higher age. If a child is not considered to be mature enough, the parents together with the school and psychologists can postpone enrollment one year. In addition, some older students return to improve their grades, and immigrants are often over-aged at graduation.

3.2 Municipalities

Norway currently has 428 municipalities located in 19 different counties. Municipalities range in size from 206 inhabitants (Utsira) to 647,676 inhabitants (Oslo). The mean and median number of inhabitants are 12,027 and 4,674 respectively (Statistics Norway, 2015). Norwegian municipalities are multipurpose institutions, providing a large number of services, such as day care and care for the elderly, in addition to primary and lower secondary education. There are usually several primary schools within each school district, but many small school districts only have one lower secondary school.

Municipality mergers

Historically, the local public sector in Norway has been divided into a large number of small municipalities and in 1957 there were more than 700 municipalities in the country. An important feature of the Norwegian system is that changes in municipality borders and splits and mergers of municipalities must be approved by the central government. Thus, the central government has always played an important role in the design of municipality structure. During the 1960's the government initiated and implemented a large merger reform reducing the number by nearly 40 percent and as a result the number of municipalities was 454 in 1982.[13]

In our empirical analysis we explore eight enforced municipality mergers occurring from 1988 to 1994 which reduced the number of municipalities from 454 to 435.[14] The municipality mergers were carried out as a result of two Official Norwegian Reports charged with recommending municipality mergers surrounding cities (Norwegian Ministry of Local Government and Labor (1986, 1989), known as Buvik I and Buvik II respectively).

The mergers in the 1960's merged many city municipalities with surrounding municipalities, but in some cases, it was argued that the mergers had not gone far enough. This was particularly true for the county of Vestfold. The city municipalities of Horten, Tønsberg and Larvik were not expanded in the 1960's and experienced problems with placement of businesses, housing, and public infrastructure generally. The city municipalities had made many attempts at merging with surrounding municipalities without success.

In the 1980's, the ministry of Local Government and Labor decided it was nec-

[13] An extensive description of the historical development of municipality structure in Norway is given in Norwegian Ministry of Local Government (1992)

[14] After 1994, there have been 7 additional voluntary mergers bringing the number of municipalities down to 428.

essary to find a solution for these city municipalities and appointed a committee to look into potential mergers in Vestfold county. The committee published the Official Norwegian Report, Norwegian Ministry of Local Government and Labor (1986), recommending specific mergers around the city municipalities of Horten, Tønsberg and Larvik. The recommended Horten merger was implemented without resistance, while the recommended mergers for Tønberg and Larvik were passed with a majority in the Parliament. All mergers were implemented January 1, 1988.

Other city municipalities with similar problems were identified while working on the Vestfold mergers, and the committee was asked to look into potential mergers for the city municipalities of Sarpsborg and Fredrikstad in the county of Østfold, Arendal in the county of Aust-Agder, Hamar in the county of Hedmark and Hammerfest in the county of Finnmark. This resulted in the second Official Norwegian Report, Norwegian Ministry of Local Government and Labor (1989). The mergers for Sarpsborg, Arendal and Hammerfest were implemented as recommended January 1, 1992 while the recommended merger for Fredrikstad was implemented as recommended January 1, 1994. As for Hamar, the recommendation was that Hamar merge with Vang, Løten and a part of Ringsaker. The resistance in Løten was so large that they were able to remain independent by a marginal vote in their favor. Hamar, Vang and parts of Ringsaker merged January 1, 1994.

The mergers were often met with large resistance by affected municipalities,[15] and in 1995 the Parliament decided municipalities should no longer be merged against their will, after which no further municipalities merged until 2002.

Table 1 shows the complete list of municipalities affected by the mergers with city municipalities in italics. In all cases, the city municipality was chosen to have the new administrative center. Although all of the mergers are city municipalities merging with surrounding municipalities, we see that the number of inhabitants in the city and surrounding municipalities are quite similar, so it is not necessarily the case that a large city is absorbing much smaller neighboring municipalities.

[15]Some municipalities organized referendums before the proposed mergers. In Onsøy, Rolvsøy, Borge, Kråkerøy, Øyestad and Vang municipality, less than 10% voted for a merger.

Table 1: Municipality mergers

Year	New municipality	Municipalities merged	Population year prior to merger
1988	Tønsberg	*Tønsberg*	8,893
		Sem	21,942
1988	Larvik	*Larvik*	8,036
		Stavern	2,538
		Tjølling	7,876
		Brunlanes	8,137
		Hedrum	10,446
1988	Horten	*Horten*	12,993
		Borre	9,095
1992	Sarpsborg	*Sarpsborg*	11,826
		Varteig	2,199
		Skjeberg	14,295
		Tune	18,288
1992	Arendal	*Arendal*	12,478
		Moland	8,148
		Øyestad	8,679
		Tromøy	4,711
		Hisøy	4,026
1992	Hamar	*Hamar*	16,351
		Vang	9,103
1992	Hammerfest	*Hammerfest*	6,909
		Sørøysund	2,341
1994	Fredrikstad	*Fredrikstad*	26,539
		Borge	11,959
		Rolvsøy	5,947
		Kråkerøy	7,445
		Onsøy	12,923

3.3 Data

The Norwegian register data from Statistics Norway cover all individuals born in 1965-1984 leaving secondary school during 1981-2000. The data contain unique identifiers that allow us combine detailed individual information including which school they attended in lower secondary school. The main outcome variables are years of education and income. Years of education is measured by degrees obtained in 2011. In higher education that is bachelor degree, master degree, and PhD, with 16, 18, and 21 years of education, respectively. Income is measured as the log of average pension qualifying income for the years 2009 and 2010. The youngest individuals are 27 years of age when education is measured and 25-26 years of age when income is measured.

The individual register data include information on gender, birth month and immigration status.[16] We also have information on parental education[17] and parental employment status[18] the year the individual turns 16, the year the individual leaves lower secondary school. Descriptive statistics are presented in Table 2.

We define the first cohort affected by the merger as the cohort leaving lower secondary school the year of the merger. As the mergers occurred January 1st, this cohort is potentially affected by the reform for half a year. All subsequent cohorts are affected for an additional year.

There are two main samples in the analysis. In the first sample, "All municipalities", merged municipalities are compared to all other municipalities in Norway. In the second sample, "Potential mergers", merged municipalities are compared to all other potential municipality mergers. These are defined as all city municipalities that existed in 1987, the year before the first merger, and all municipalities bordering the city municipalities within the same county.[19] The sample includes 211 municipalities (46% of all municipalities) displayed in Figure 1. For both samples, the sample of merged municipalities includes a window of 10+/- years around the merger year. This time period is shortened for each merger either due to data only being available from 1981 or due to the data ending in 2000. All available years are included for the non-merged municipalities.

We restrict the sample to students turning 16 the year they graduate from lower secondary school. The cohort leaving school in 1990 has missing information on school identifies, and is therefore not included in the analysis. Students with missing information on income or years of education are excluded from the analysis. Table A1 reports the observations lost due to these restrictions in the "All municipalities" and in the "Potential mergers" sample.

[16]Immigration status is divided into first and second generation immigrant, where first generation immigrants are born abroad and have both parents born abroad, while second generation immigrants are born in Norway and have both parents born abroad.

[17]Parental education is categorized as the highest completed education by one of the parents. The categories included are upper secondary education (High school), Bachelor's degree, Master's degree or PhD, and unknown education, with less than upper secondary education being the reference category.

[18]Indicators for only mother working, only father working, and both parents working are included, with the reference category being no parent working.

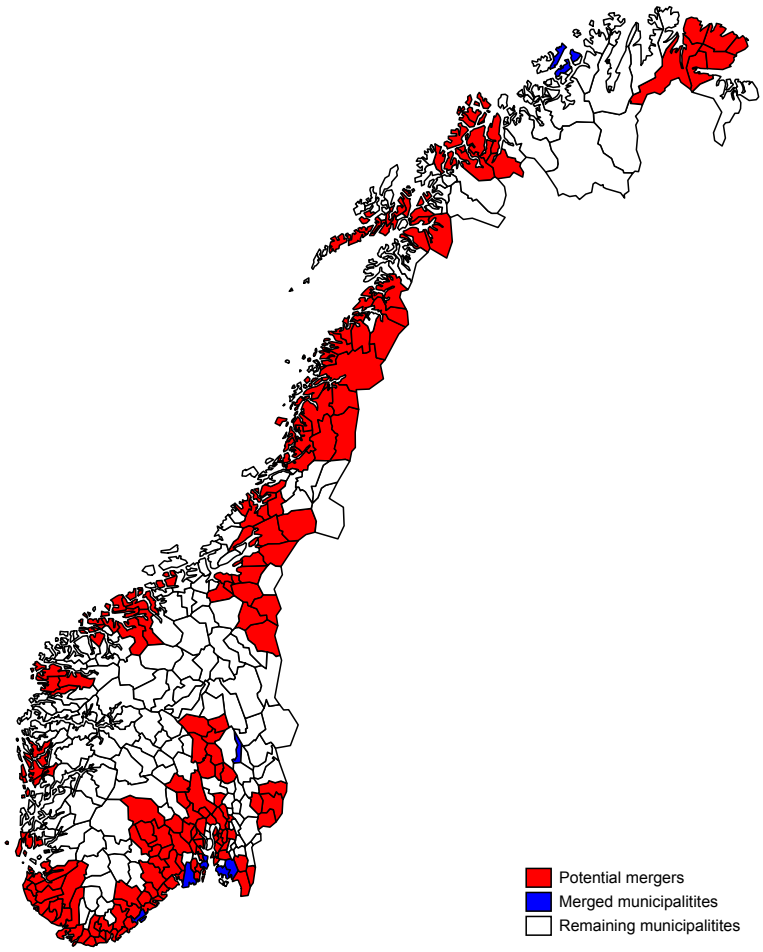
[19]For Oslo, all bordering municipalities are included regardless of county since Oslo is both a municipality and a county.

Table 2: Descriptive statistics

	Treated			Comparison all municipalities			Comparison potential mergers		
	mean	(sd)	N	mean	(sd)	N	mean	(sd)	N
A. Outcome variables									
Log of income 2009-2010	12.7	(0.75)	56245	12.7	(0.77)	924876	12.7	(0.79)	668313
Years of education	14	(2.55)	59635	13.9	(2.54)	976519	14	(2.57)	707819
B. Socioeconomic characteristics									
Girl	0.49	(0.50)	59635	0.49	(0.50)	976519	0.49	(0.50)	707819
Parental education: High School	0.56	(0.50)	59635	0.54	(0.50)	976519	0.53	(0.50)	707819
Parental education: Bachelor	0.21	(0.40)	59635	0.2	(0.40)	976519	0.2	(0.40)	707819
Parental education: Masters +	0.066	(0.25)	59635	0.077	(0.27)	976519	0.085	(0.28)	707819
Parental education: Unknown	0.028	(0.16)	59635	0.032	(0.18)	976519	0.034	(0.18)	707819
First generation immigrant	0.009	(0.09)	59635	0.013	(0.11)	976519	0.015	(0.12)	707819
Second generation immigrant	0.004	(0.06)	59635	0.006	(0.08)	976519	0.008	(0.09)	707819
Only mother working	0.17	(0.37)	59635	0.17	(0.37)	976519	0.17	(0.37)	707819
Only father working	0.16	(0.37)	59635	0.15	(0.35)	976519	0.15	(0.36)	707819
Both parents working	0.31	(0.46)	59635	0.33	(0.47)	976519	0.33	(0.47)	707819
Birth month	6.26	(3.33)	59635	6.35	(3.33)	976519	6.35	(3.33)	707819
C. Municipality characteristics (log)									
Total population	10.3	(0.55)	136	8.45	(1.02)	8001	8.8	(1.11)	3921
School aged population	8.15	(0.54)	136	6.39	(1.03)	8001	6.75	(1.11)	3921
16-year olds	5.95	(0.56)	136	4.17	(1.04)	8001	4.53	(1.11)	3921
Total expenditures	20.2	(0.53)	136	18.5	(0.90)	7999	18.8	(1.03)	3920
Per capita total expenditures	9.85	(0.28)	136	10	(0.39)	7999	9.97	(0.38)	3920
School expenditures	18.8	(0.47)	136	17.3	(0.88)	8000	17.6	(0.96)	3920
Per student school expenditures	10.7	(0.20)	136	10.9	(0.28)	8000	10.8	(0.27)	3920
Teachers without teacher certification	1.68	(0.94)	127	1.41	(0.94)	6825	1.55	(1.01)	3329
Lower secondary schools	1.66	(0.40)	136	0.55	(0.64)	7981	0.72	(0.73)	3903

Note: Descriptive statistics corresponding to the estimation sample for years of education. Treated includes all individuals from municipalities experiencing a merger. Comparison all municipalities includes all non-merged municipalities. Comparison potential mergers includes all non-merged city municipalities and their bordering municipalities in 1987. All municipality characteristics are measured in log. Errors in reporting school and total expenditures reduce N for these variables. For teachers without teacher certification and lower secondary schools, N is reduced due to observations with 0.

Figure 1: Potential mergers and merged municipalities



4 Identification and model specification

The merges are investigated using a difference-in-differences model estimated by OLS. *Treat* is equal to one if the individual graduated from a lower secondary school located in a municipality that merged sometime between 1981 and 2000. This includes all municipalities in Table 1. *Post* is equal to one in the time period after the merger for the cohorts thought to be affected by the merger. α_t is a cohort specific constant term and corresponds to age at graduation as we restrict our sample to students graduating from lower secondary school the year they turn 16. The cohort specific constant term consumes the separate effect of the variable *Post*.

This model can be expressed as

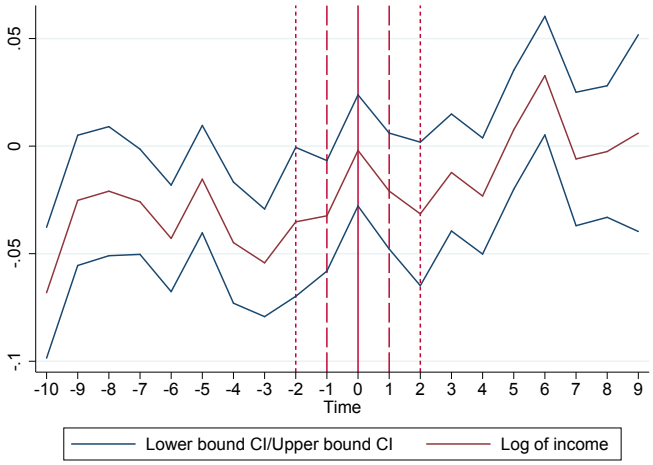
$$Y_{it} = \alpha_t + \beta Treat_i + \gamma Treat_i \times Post_t + X_{it}^{'}\delta + \epsilon_{it} \tag{1}$$

where i indexes individual and t indexes cohort. X indicates the socioeconomic characteristics of the individual, and includes individual characteristics (immigrant status, gender and birth month) and parental characteristics (parental education and employment status). Socioeconomic characteristics are measured the year the individual leaves lower secondary school. Standard errors, ϵ_{it}, are clustered at the school level.

We want to compare the outcomes of students in treated municipalities before and after the merger to students in non-treated municipalities before and after the merger. $Treat_i \times Post_t$ is our variable of interest, and γ captures this effect. If the change in outcomes from the pre-merger period to the post-merger period is significantly different in the merged municipalities than in the non-merged municipalities, then γ will be significantly different from 0. If γ is significant and positive, this indicates that the merger has a positive effect on outcomes and the opposite if γ is significant and negative.

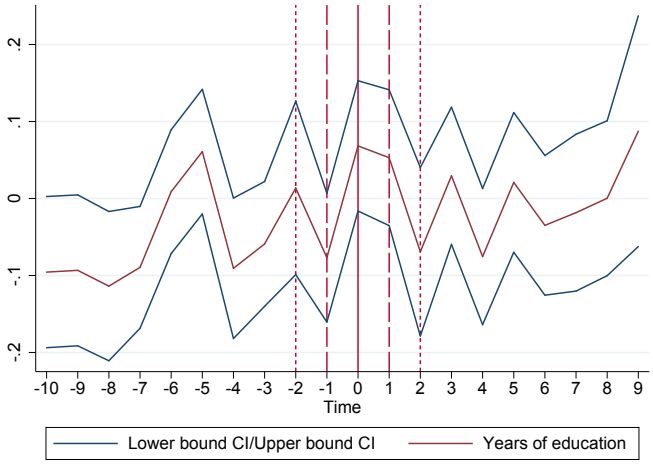
Figures 2 and 3 investigate whether the parallel trends assumption holds. The treatment (the mergers) occurred in different years in different municipalities. The figures present log of income (Figure 2) and years of education (Figure 3) relative to the control municipalities in the "Potential mergers" sample. Log of income and years of education in treated municipalities are compared to the non-treated municipalities in the same year for each individual observation. The red lines present the mean values while the blue lines present the 95% confidence interval. Time indicates the time period relative to the treatment year where the treatment year is time=0.

Figure 2: Trend in the relative log of income



Note: Trend in the relative log of income with 95% confidence interval. Time indicates time relative to treatment year, with 0 being the first year of treatment (solid red line). In Table 4 observations between the long dashed lines are dropped from estimations in column (2) and observations between the short dashed lines are dropped from estimations in column (4).

Figure 3: Trend in the relative years of education



Note: Trend in the relative years of education with 95% confidence interval. Time indicates time relative to treatment year, with 0 being the first year of treatment (solid red line). In Table 4 observations between the long dashed lines are dropped from estimations in column (2) and observations between the short dashed lines are dropped from estimations in column (4).

Both figures show some variation in the relative measures. However, the figures do not show a clear pre-treatment trend, which supports the parallel trends assumption. Relative log of income increases after the mergers indicating that income is increasing in treated municipalities relative to non-treated municipalities after the

merger. The pattern is not as clear for relative years of education, and it is unclear whether the mergers increased years of education.

5 Results

We run three versions of Equation (1). In the first, we exclude socioeconomic characteristics, in the second we include socioeconomic characteristics, and in the third version we add school fixed effects. Adding school fixed effects allows us to control for time-invariant unobserved differences between individuals from different schools. Results with the sample "All municipalities" are presented in columns (1)-(3) of Table 3. Results with the sample "Potential mergers" are presented in columns (4)-(6). The top panel displays results for log income while the bottom panel displays results for years of education.

For log income, estimates show that municipality mergers have a positive effect on income. After the merger, income increases by about 2-3 % in the merged municipalities compared to the non-merged municipalities. With the "All municipalities" sample, the effect is approximately 2%. The effect increases to 3% when the comparison group consists of the sample "Potential mergers". For years of education, the estimates are positive for years of education (about 0.05), but they are not significant at conventional levels when including school fixed effects. This is true for both samples, where the t-value is equal to 1.6 in the "All municipalities" sample and 1.3 in the "Potential mergers" sample.

Both samples confirm the same results. We believe the "Potential mergers" sample to be the best suited for this difference-in-differences specification. In Sections 6 and 7, estimates are reported using the "Potential mergers" sample along with time/age fixed effects, socioeconomic characteristics and school fixed effects.

Table 3: Effect of mergers on log income and years of education

	All municipalities			Potential mergers		
	(1)	(2)	(3)	(4)	(5)	(6)
A. Dependent variable: Log income						
Treat*Post	0.0205**	0.0182*	0.0206**	0.0310***	0.0269**	0.0298***
	(0.0102)	(0.0103)	(0.0102)	(0.0104)	(0.0104)	(0.0103)
Treat	-0.0327***	-0.0279***		-0.0360***	-0.0291***	
	(0.0075)	(0.0063)		(0.0077)	(0.0064)	
Observations	981,126	981,126	981,126	724,561	724,561	724,561
R-squared	0.049	0.107	0.106	0.051	0.108	0.106
No. of schools			1,402			920
B. Dependent variable: Years of education						
Treat*Post	0.0514	0.0535*	0.0461	0.0572*	0.0500	0.0417
	(0.0325)	(0.0303)	(0.0316)	(0.0332)	(0.0308)	(0.0320)
Treat	-0.0214	-0.00433		-0.0518	0.0104	
	(0.0575)	(0.0329)		(0.0590)	(0.0336)	
Observations	1,036,154	1,036,154	1,036,154	767,454	767,454	767,454
R-squared	0.007	0.168	0.150	0.007	0.170	0.150
No. of schools			1,413			929
Time/age FE	Yes	Yes	Yes	Yes	Yes	Yes
Soc. Char.	No	Yes	Yes	No	Yes	Yes
School FE	No	No	Yes	No	No	Yes

Note: Standard errors clustered at the school level in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Socioeconomic characteristics include birth month, gender, immigration status, parental education, and parental employment status.

6 Robustness checks

This section presents results for various model specifications. The results are presented in Table 4 and all should be compared to column (6) of Table 3.

First we investigate whether results are sensitive to excluding the two biggest cities from our sample, Bergen and Oslo. Results are reported in column (1) of Table 4. Both cities, along with their bordering municipalities are excluded from the estimation which reduces the sample by 26%. The estimate for log income is reduced from 3% to 2% but is still significant. For years of education, the estimate increases from 0.04 to 0.06 and is significant at the 10% level (t-value of 1.8).

Next, the years right before and after the merger are removed from the estima-

tion, creating a "donut hole". The first cohort affected by the merger is only in school for 6 months after the merger. This might not be sufficient time to expect there to be an effect. Also, there could be some anticipatory effects of the merger which would affect the cohorts leaving lower secondary school just before the merger. Removing the observations just around the time of the merger removes such concerns.

Column (2) reports the results when removing the one observation before and one after (time= -1 and time= 0). Column (3) reports results when two years are removed before and after the merger (time= -2 and time= 1 are also removed). The long dashed lines in Figures 2 and 3 correspond to the 2-year "donut hole" (column (2)) while the short dashed lines correspond to the 4-year "donut hole" (column (3)). In both specifications the results remain strongly significant for log income. The estimate is 3% for the 2-year "donut hole" and 3.5% for the 4-year "donut hole". For education, results are insignificant.

Another concern is the length of our estimation window. In our main results, the estimation window is 10 years before and after the reform (when possible). Column (4) estimates the results when reducing this window to 5 years. This reduces the point estimate to 1.7% for income with a t-value of 1.7. For each year following the merger, the cohort leaving lower secondary school has spent an additional year in a post-merge school. If there is an effect of the merger through schools, then we would expect this effect to be larger for later cohorts. It is therefore expected that this estimate is somewhat lower. For years of education, the results are very similar to column (6) of Table 3.

Finally, we run a placebo reform. In this specification, we pretend that the merger happened 4 years before and only include pre-merger years for the treated municipalities. A significant estimate in this specification would challenge our common trends assumption. For both log income and years of education, estimates are insignificant. The estimate is -1.2% for income with a t-value of 0,984. For years of education, the estimate is -0.04 with a t-value of 1.27.

Table 4: Robustness checks

	(1) No big cities	(2) 2 year «donut hole»	(3) 4 year «donut hole»	(4) 5-year window	(5) Placebo reform
A. Dependent variable: Log income					
Treat*Post	0.0203**	0.0309***	0.0345***	0.0166*	-0.0123
	(0.0103)	(0.0114)	(0.0124)	(0.0098)	(0.0125)
Observations	537,513	717,809	712,854	561,246	465,775
R-squared	0.108	0.106	0.106	0.088	0.086
No. of schools	696	920	920	889	870
B. Dependent variable: Years of education					
Treat*Post	0.0585*	0.0241	0.00954	0.0451	-0.0423
	(0.0326)	(0.0347)	(0.0358)	(0.0316)	(0.0332)
Observations	567,327	760,317	755,087	594,036	497,141
R-squared	0.150	0.150	0.150	0.150	0.145
No. of schools	703	929	929	896	873

Note: All regressions include time/age fixed effects, socioeconomic characteristics and school fixed effects. No big cities drops the city municipalities Oslo and Bergen along with their bordering municipalities. 2 and 4 year "donut hole" drop the 1+/- and 2+/- years surrounding the merger. 5-year window reduces the estimation window to 5 +/- years surrounding the merger. Placebo reform runs the specification as if the merger occurred 4 years earlier and only includes years before the merger occurred. The sample corresponds to the "Potential mergers" sample. Standard errors clustered at the school level in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Socioeconomic characteristics include birth month, gender, immigration status, parental education, and parental employment status.

7 Mechanisms

What are the mechanisms behind the main results? First, we investigate if results differ depending on whether the student attends a school in a city or a surrounding municipality. Next, we investigate whether municipality characteristics change before and after the merger using the (merged) municipality as the unit of analysis.

7.1 City vs. surrounding schools

A unique feature of our data set is that we can separate between city municipality schools and surrounding municipality schools both before and after the merger. This allows us to study the effect for students attending city school and surrounding schools separately.

Table 5 displays the results. Column (1) the same specification as column (6) of Table 3. In column (2), only students from city schools are included in the analysis. This includes students in city municipalities that experience a merger and students in city municipalities in the "Potential mergers" comparison group. The point estimate for log income is small (0.7%) and the results are nowhere close to being significant. The point estimate for years of education is negative and not significant.

In column (3), only students from surrounding schools are included in the analysis. This includes students in surrounding municipalities that experience a merger and students in surrounding municipalities in the "Potential mergers" comparison group. The point estimate for log income is 3%, and is highly significant, while the point estimate for years of education is 0.06 and not significant at conventional levels (t-value of 1.58). This shows that the results are driven by students from surrounding schools.

Table 5: Mechanisms – city vs. surrounding schools

	(1) All schools	(2) City schools	(3) Surrounding schools
Dependent variable: Log income			
Treat*Post	0.0298***	0.00710	0.0293***
	(0.0103)	(0.0181)	(0.0110)
Observations	724,561	410,248	314,421
R-squared	0.106	0.105	0.108
No. of schools	920	462	461
Dependent variable: Years of education			
Treat*Post	0.0417	-0.0448	0.0616
	(0.0320)	(0.0414)	(0.0391)
Observations	767,454	436,567	331,013
R-squared	0.150	0.151	0.149
No. of schools	929	469	462

Note: All regressions include time/age fixed effects, socioeconomic characteristics and school fixed effects. The sample corresponds to the "Potential mergers" sample. Standard errors clustered at the school level in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Socioeconomic characteristics include birth month, gender, immigration status, parental education, and parental employment status.

7.2 Municipality effects

Lastly, we investigate whether municipality characteristics change in merged municipalities relative to non-merged municipalities after the merger. Total population

and school aged population are from Statistics Norway while the number of schools and 16-year olds are constructed from our data. Expenditure measures are from municipality accounts and the share of certified teachers is a measure previously used by Bonesrønning, Falch, and Strøm (2005) and Falch, Johansen, and Strøm (2009).

Table 6 displays results where estimations include one observation per (merged) municipality and year. In Column (1), the outcome is the log of total population in the municipality. Columns (2) and (3) include the school aged population and the 16 year olds respectively. All estimates are insignificant. There is no evidence of demographic changes resulting from the mergers.

Table 6: Mechanisms – Municipality characteristics, population

	(1)	(2)	(3)
	Total population (log)	School aged population (log)	16 year-olds (log)
Treat*Post	0.00631	-0.00249	-0.0654
	(0.0365)	(0.0434)	(0.0476)
Treat	1.507***	1.407***	1.454***
	(0.2192)	(0.2128)	(0.2065)
Observations	4,057	4,057	4,057
R-squared	0.058	0.056	0.060
Time FE	Yes	Yes	Yes

Note: The sample corresponds to the "Potential mergers" sample. The estimation includes one observation per (merged) municipality and year. Standard errors clustered at the municipality level in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

In Table 7, the first four columns display outcomes related to expenditures, measured in log. The first is total expenditures, the second is per capita total expenditures, the third is school expenditures and the fourth is per student school expenditures. The only significant estimate is the per capita total expenditures where merged municipalities have 3.8% lower expenditures after the merger compared to non-merged municipalities. This result is qualitatively consistent with the evidence in Reingewertz (2012) although numerically smaller. The effect on expenditure per student (6-15 years old) is also negative but not statistically significant. This suggests that the positive student income effect in adulthood cannot be explained by increased total budgets in merged municipalities or budget reallocation in favor of the education sector.

Table 7: Mechanisms – Municipality characteristics, population

	(1) Total exp. (log)	(2) Per capita total exp. (log)	(3) School exp. (log)	(4) Per student school exp. (log)	(5) Teachers w/o teacher certification (log)	(6) Lower secondary schools (log)
Treat*Post	-0.0317	-0.0384**	-0.0274	-0.0262	0.0273	-0.0281
	(0.0314)	(0.0169)	(0.0351)	(0.0261)	(0.2392)	(0.0312)
Treat	1.405***	-0.102*	1.276***	-0.131**	0.143	0.943***
	(0.1759)	(0.0536)	(0.1741)	(0.0543)	(0.2761)	(0.1549)
N	4,056	4,056	4,056	4,056	3,456	4,039
R-squared	0.134	0.537	0.068	0.254	0.048	0.053
Time FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: The sample corresponds to the "Potential mergers" sample. The estimation includes one observation per (merged) municipality and year. Errors in reporting school and total expenditures reduce N compared to Table 7 for these variables. For teachers without teacher certification and lower secondary schools, N is reduced due to observations with 0. Standard errors clustered at the municipality level in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

In column (8), the outcome is teachers without teacher certifications. The variable teachers without teacher certification is a reasonable measure of teacher quality (Bonesrønning, Falch, and Strøm, 2005; Falch, Johansen, and Strøm, 2009). Our results do not seem to be driven by increased teacher quality. The last column is the number of lower secondary schools in the municipality. There is no evidence of a change in the number of lower secondary schools as a result of the merger.

8 Conclusion

We use the spatial and temporal variation in municipality merges in a difference-in-differences approach to provide quasi-experimental evidence of the effect of municipality size on school output measured by student educational attainment and income in adulthood. Municipality mergers are found to increase student income in adulthood by 2-3%, while the effect on educational attainment is generally positive, but not so precisely estimated.

Our results are consistent with the hypothesis that student enrolled in schools in former surrounding municipalities took advantage of potential gains in existing administrative quality in the former cities. The income effect is driven by students enrolled in schools in premerger municipalities surrounding the former city, not by students enrolled in premerger city schools. However, further research is needed to confirm this interpretation.

We also find that the merger reduced total municipal expenditure per capita by

nearly 5% which is qualitatively consistent with the evidence in Reingewertz (2012) although numerically smaller. The effect on expenditure per student (6-15 years old) is also negative but not statistically significant. This suggests that the positive student income effect in adulthood cannot be explained by increased total budgets in merged municipalities or budget reallocation in favor of the education sector. Finally, we find that the number of lower secondary schools, the number of persons aged 7-16 and overall teacher quality measured by the share of teachers without a teacher certification at the municipality level is not significantly affected by the merger. Thus, we tentatively conclude that systematic changes in the number of schools, cohort size and teacher quality cannot explain the income effect.

When deciding whether to merge municipalities, proponents argue that larger municipalities increase efficiency, while opponents argue that the population is further removed from their elective officials. The results from this paper suggest that municipality mergers can have positive effects on school outputs measured by years of education and income in adulthood, lending support to the proponents of municipality mergers.

References

Aakvik, A., K. G. Salvanes, and K. Vaage (2010): "Measuring heterogeneity in the returns to education using an education reform," *European Economic Review*, 54(4), 483–500.

Abdulkadiroğlu, A., W. Hu, and P. A. Pathak (2013): "Small high schools and student achievement: lottery-based evidence from New York City," Discussion paper, National Bureau of Economic Research.

Alesina, A., R. Baqir, and C. Hoxby (2004): "Political Jurisdictions in Heterogeneous Communities," *Journal of Political Economy*, 112(2), 348–396.

Alesina, A., and E. Spolaore (1997): "On the Number and Size of Nations," *Quarterly Journal of Economics*, 112(4), 1027–1056.

Andrews, M., W. Duncombe, and J. Yinger (2002): "Revisiting economies of size in American education: are we any closer to a consensus?," *Economics of Education Review*, 21(3), 245–262.

Angrist, J., E. Battistin, and D. Vuri (2015): "In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno," IZA Discussion Papers 8959, Institute for the Study of Labor (IZA).

Atkinson, A. B., and J. E. Stiglitz (1980): *Lectures on public economics*. Maidenhead: McGraw-Hill.

Barankay, I., and B. Lockwood (2007): "Decentralization and the productive efficiency of government: Evidence from Swiss cantons," *Journal of Public Economics*, 91(5), 1197–1218.

Barrow, L., D. W. Schanzenbach, and A. Claessens (2015): "The impact of Chicago's small high school initiative," *Journal of Urban Economics*, 87, 100–113.

Baum, D. N. (1986): "A simultaneous equations model of the demand for and production of local public services: The case of education," *Public Finance Review*, 14(2), 157–178.

Berry, C. R., and M. R. West (2010): "Growing pains: The school consolidation movement and student outcomes," *Journal of Law, Economics, and Organization*, 26(1), 1–29.

Besley, T., and A. Case (1995): "Incumbent behavior: vote seeking, tax setting and yardstick competition," *American Economic Review*, 85(1), 25–45.

Besley, T., and S. Coate (2003): "Centralized versus decentralized provision of local public goods: a political economy approach," *Journal of Public Economics*, 87(12), 2611–2637.

Beuchert, L. V., M. K. Humlum, H. S. Nielsen, and N. Smith (2015): "The Short-Term Effects of School Consolidation on Student Achievement: Evidence of Disruption?," *Available at SSRN 2626712*.

Bonesrønning, H., T. Falch, and B. Strøm (2005): "Teacher sorting, teacher quality, and student composition," *European Economic Review*, 49(2), 457–483.

Borge, L.-E., J. K. Brueckner, and J. Rattsø (2014): "Partial fiscal decentralization and demand responsiveness of the local public sector: Theory and evidence from Norway," *Journal of Urban Economics*, 80, 153–163.

Brennan, G., and J. M. Buchanan (1980): *The power to tax: Analytic foundations of a fiscal constitution*. Cambridge University Press.

Breunig, R., and Y. Rocaboy (2008): "Per-capita public expenditures and population size: a non-parametric analysis using French data," *Public Choice*, 136(3-4), 429–445.

Brummet, Q. (2014): "The effect of school closings on student achievement," *Journal of Public Economics*, 119, 108–124.

Callan, S. J., and J. M. Thomas (2001): "Economies of scale and scope: A cost analysis of municipal solid waste services," *Land Economics*, 77(4), 548–560.

Coate, S., and B. Knight (2007): "Socially optimal districting: a theoretical and empirical exploration," *Quarterly Journal of Economics*, pp. 1409–1471.

Dafflon, B. (2013): "Voluntary amalgamation of local governments: the Swiss debate in the European context," in *The Challenge of Local Government Size: Theoretical Perspectives, International Experience and Policy Reform*. Edward Elgar Publishing.

de Haan, M., E. Leuven, and H. Oosterbeek (2014): "School supply and student achievement: Evidence from a school consolidation reform," *Mimeo*, 1, 2014.

DeBoer, L. (1992): "Economies of scale and input substitution in public libraries," *Journal of Urban Economics*, 32(2), 257–268.

Derksen, W. (1988): "Municipal amalgamation and the doubtful relation between size and performance," *Local Government Studies*, 14(6), 31–47.

Driscoll, D., D. Halcoussis, and S. Svorny (2003): "School district size and student performance," *economics of Education Review*, 22(2), 193–201.

Duncombe, W., and J. Yinger (2007): "Does school district consolidation cut costs?," *Education Finance and Policy*, 2(4), 341–375.

Ellingsen, T. (1998): "Externalities vs internalities: a model of political integration," *Journal of Public Economics*, 68(2), 251–268.

Engberg, J., B. Gill, G. Zamarro, and R. Zimmer (2012): "Closing schools in a shrinking district: Do student outcomes depend on which schools are closed?," *Journal of Urban Economics*, 71(2), 189–203.

Falch, T., K. Johansen, and B. Strøm (2009): "Teacher shortages and the business cycle," *Labour Economics*, 16(6), 648–658.

Farsi, M., A. Fetz, and M. Filippini (2007): "Economies of scale and scope in local public transportation," *Journal of Transport Economics and Policy*, pp. 345–361.

Ferguson, R. F. (1991): "Paying for public education: New evidence on how and why money matters," *Harvard Journal of Legislation*, 28, 466–498.

Ferguson, R. F., and H. F. Ladd (1996): "Additional evidence on how and why money matters: A production function analysis of Alabama schools," in *Holding schools accountable: Performance-based reform in education*, ed. by H. F. Ladd. Washington, DC: Brookings Institution Press.

Fox, W. F. (1981): "Reviewing economies of size in education," *Journal of Education Finance*, pp. 273–296.

Gordon, N., and B. Knight (2008): "The effects of school district consolidation on educational cost and quality," *Public Finance Review*, 36(4), 408–430.

Gyimah-Brempong, K. (1987): "Economies of scale in municipal police departments: The case of Florida," *Review of Economics and Statistics*, pp. 352–356.

Hansen, S. W. (2014): "Common pool size and project size: an empirical test on expenditures using Danish municipal mergers," *Public Choice*, 159(1-2), 3–21.

Heinesen, E. (2005): "School district size and student educational attainment: evidence from Denmark," *Economics of Education Review*, 24(6), 677–689.

Hinnerich, B. T. (2009): "Do merging local governments free ride on their counterparts when facing boundary reform?," *Journal of Public Economics*, 93(5), 721–728.

Humlum, M. K., and N. Smith (2015): "Long-term effects of school size on students' outcomes," *Economics of Education Review*, 45, 28–43.

Jacques, C., B. W. Brorsen, and F. G. Richter (2000): "Consolidating rural school districts: Potential savings and effects on student achievement," *Journal of Agricultural and Applied Economics*, 32(03), 573–583.

Jordahl, H., and C.-Y. Liang (2010): "Merged municipalities, higher debt: on free-riding and the common pool problem in politics," *Public Choice*, 143(1-2), 157–172.

Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014): "Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success," OECD Education Working Papers 110, OECD Publishing.

Kiesling, H. J. (1967): "Measuring a local government service: A study of school districts in New York State," *Review of Economics and Statistics*, pp. 356–367.

Kraus, M. (1981): "Scale economies analysis for urban highway networks," *Journal of Urban Economics*, 9(1), 1–22.

Kuziemko, I. (2006): "Using shocks to school enrollment to estimate the effect of school size on student achievement," *Economics of Education Review*, 25(1), 63–75.

Liu, C., L. Zhang, R. Luo, S. Rozelle, and P. Loyalka (2010): "The effect of primary school mergers on academic performance of students in rural China," *International Journal of Educational Development*, 30(6), 570–585.

Lockwood, B. (2002): "Distributive Politics and the Costs of Centralization," *Review of Economic Studies*, 69(2), 313–337.

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

Moisio, A., and R. Uusitalo (2013): "The impact of municipal mergers on local public expenditures in Finland," *Public Finance and Management*, 13(3), 148.

Musgrave, R. A., and P. B. Musgrave (1973): *Public finance in theory and practice (2nd ed)*. Tokyo: McGraw-Hill.

Niskanen, W. A. (1998): "Student performance and school district size," in *Policy Analysis and Public Choice*, ed. by W. A. Niskanen. Cheltenham, UK: Edward Elgar.

Norwegian Ministry of Local Government (1992): *NOU 1992:15: Kommune- og fylkesinndelingen i et Norge i forandring (Municipality and county division in a changing Norway).*

Norwegian Ministry of Local Government and Labor (1986): *NOU 1986:7: Forslag til endringer i kommuneinndelingen for byområdene Horten, Tænsberg og Larvik i Vestfold fylke (Suggestions to changes in the municipality structure for the city areas Horten, Tænsberg and Larvik in Vestfold county).*

——— (1989): *NOU 1989:16: Kommuneinndelingen for byområdene Sarpsborg, Fredrikstad, Arendal, Hamar og Hammerfest (Municipality structure for the city areas Sarpsborg, Fredrikstad, Arendal, Hamar and Hammerfest).*

Oates, W. E. (1972): *Fiscal Federalism*. New York: Harcourt Brace.

——— (1985): "Searching for Leviathan: An empirical study," *American Economic Review*, 75(4), 748–757.

——— (2005): "Toward a second-generation theory of fiscal federalism," *International Tax and Public Finance*, 12(4), 349–373.

Razin, A. (1999): "Budget differences between large and small municipalities in Israel," *Floersheimer Institute of Policy Studies, Jerusalem*.

Reingewertz, Y. (2012): "Do municipal amalgamations work? Evidence from municipalities in Israel," *Journal of Urban Economics*, 72(2), 240–251.

Reiter, M., and A. Weichenrieder (1997): "Are public goods public? A critical survey of the demand estimates for local public services," *FinanzArchiv/Public Finance Analysis*, pp. 374–408.

Saarimaa, T., and J. Tukiainen (2015): "Common pool problems in voluntary municipal mergers," *European Journal of Political Economy*, 38, 140–152.

Schwartz, A. E., L. Stiefel, and M. Wiswall (2013): "Do small schools improve performance in large, urban districts? Causal evidence from New York City," *Journal of Urban Economics*, 77, 27–40.

Sebold, F. D., and W. Dato (1981): "School funding and student achievement: An empirical analysis," *Public Finance Review*, 9(1), 91–105.

Solé-Ollé, A., and N. Bosch (2005): "On the relationship between authority size and the costs of providing local services: lessons for the design of intergovernmental transfers in Spain," *Public Finance Review*, 33(3), 343–384.

Statistics Norway (2015): "StatBank Norway," `http://www.ssb.no/en/statistikkbanken`, Reading date: 15.12.2015.

Tiebout, C. M. (1956): "A pure theory of local expenditures," *Journal of Political Economy*, pp. 416–424.

Walberg, H. J., and W. J. Fowler (1987): "Expenditure and size efficiencies of public school districts," *Educational Researcher*, 16(7), 5–13.

Zax, J. S. (1989): "Is there a Leviathan in your neighborhood?," *American Economic Review*, 79(3), 560–567.

A Appendix

Table A1: Data reduction

	All municipalities		Potential mergers	
	Observations	% Reduc.	Observations	% Reduc.
1. Sample 1982-2000 (without 1990)	1105383		823700	
2. Non-missing municipality	1103880	0,14 %	822197	0,18 %
3. 16 years old when graduating from lower secondary school	1044816	5,35 %	775671	5,66 %
4. 10 +/- years around merge	1036919	0,76 %	768072	0,98 %
5. Non missing years of education	1036154	0,07 %	767454	0,08 %
5. Non missing log of income	981126	5,38 %	724561	5,67 %

Note: Data on the school identifier is missing in 1990. 55,789 and 43,508 observations have zero income for all municipalities and potential mergers respectively. They excluded from the analysis because we use the logarithmic value of income.

Chapter 4:

Grade variance

Astrid Marie Jorde Sandsør

Grade variance*

Astrid Marie Jorde Sandsør†

Abstract

This paper investigates the importance of the second moment of individual grade distribution; grade variance. Transcript data from the U.S. National Longitudinal Survey of Youth, 1979, along with detailed register information for students in Norway are used to investigate the association between grade variance and educational attainment. For both the United States and Norway, grade variance is negatively associated with educational attainment across the grade distribution. Estimates are robust to controlling for socioeconomic characteristics and school fixed effects and remain negative for both genders and when including measures of cognitive and non-cognitive skills. My results suggest that institutions should consider more than just grade point average in admission decisions.

Keywords: grades, cognitive skills, non-cognitive skills, human capital
JEL codes: I21, J24

*Thanks to Angela Duckworth and the Character Development in Adolescence Project as well as Torberg Falch and the project Governance, Management and Performance in the Norwegian Educational System financed by the Norwegian Research Council (grant no. 197760) for allowing me to use their data. Thank you to Torberg Falch, Kalle Moene and Alexander Koch as well seminar participants at the Norwegian University of Science and Technology and the Workshop on Economics of Education, Mainz, September 2015 for helpful comments and suggestions. The paper is part of the research activities at the ESOP center at the Department of Economics, University of Oslo. ESOP is supported by The Research Council of Norway (S/179552).

†Department of Economics, University of Oslo

1 Introduction

What are the effects of the individual distribution of skills on school attainment and school performance? We know that cognitive skills are an important predictor for future outcomes for the individual, including education and labor market outcomes (Murnane, Willett, and Levy, 1995; Herrnstein and Murray, 2010; Heckman, 1995), and aggregate measures of cognitive skills are important for economic growth and development (Hanushek and Woessmann, 2008; Hanushek and Kimko, 2000). However, for a given average level of skills, is it better that skills are evenly divided between subject areas or is it better to be particularly good at some subject area?

One measure of cognitive skills is student grades received in school, commonly measured as the grade point average. Grades are highly correlated with short-term and long-term outcomes such as educational attainment and income. Additionally, grades have direct consequences for students, by for instance forming part of the college admission decision and determining their post-education job qualifications. Grade point average captures the first moment of the individual grade distribution, the mean. The second moment of the distribution, the variance, is a measure of grade dispersion; how far the grades are from the individual's mean. For a given grade point average, which student might be expected to have higher educational attainment; the student with high or low grade variance?

On the one hand, grades might reflect non-cognitive skills, such as motivation, perseverance and conscientiousness which have been shown to be meaningful predictors of educational, labor market and behavioral outcomes. If high grade variance is associated with low non-cognitive skills and vice versa, then a negative relationship between grade variance and educational attainment is expected. On the other hand, grades might mainly reflect knowledge in the subject, i.e., cognitive skills. As higher education allows students to specialize in their preferred field, high variance students, who are particularly good in some subjects, might be expected to have a higher educational attainment.

As there are reasons to believe that grade variance could be either positively or negatively associated with educational attainment, this makes grade variance particularly interesting to study empirically. Finding a negative association between grade variance and educational attainment, especially at the lower end of the grading distribution, supports the non-cognitive skills hypothesis while finding a positive association, especially at the upper end of the grading distribution, supports the generalist/specialist hypothesis.

In order to investigate the importance of grade variance empirically, I use three different data sources; The U.S. National Longitudinal Survey of Youth, 1979 (NLSY79), Norwegian register data (NRD) and data from the Character Development in Adolescence Project (CDAP). The NLSY79 is a longitudinal survey with a nationally representative sample of young Americans first interviewed in 1979 and includes high school transcript data, educational attainment and socioeconomic characteristics. The NRD contains the entire population of students graduating from lower secondary education in Norway from

2002-2004 and includes transcript data, educational attainment and socioeconomic characteristics. The CDAP is a longitudinal survey of middle school students and their teachers from 8 different schools and includes transcript data along with various self-reported and teacher-reported measures of non-cognitive skills.

The NLSY79 and NRD are both used to investigate the association between grade variance and educational attainment and whether the association differs across the grading distribution or by gender. The NLSY79 includes long-run educational outcomes while the NRD only includes short-run educational outcomes. In Norway, grades are the main determinant of acceptance into upper secondary and higher education, and grading practices are monitored by central authorities, reducing potential measurement error. Along with the richness of register data, this allows for a more detailed analysis in the NRD than in NLSY79. By investigating data from two different countries, I am able to investigate whether the association between grade variance is context specific or more general.

Next, the paper investigates how grade variance is associated with cognitive and non-cognitive skills. The NLSY79 includes measures of cognitive and non-cognitive skills previously used by Heckman, Stixrud, and Urzua (2006) while a subset of grades is used as measures of cognitive and non-cognitive skills in the NRD. However, in both data sets the measures of cognitive and non-cognitive are simple and may not be capturing the skills that could be expected to be associated with grade variance. The CDAP includes grades together with a rich set of non-cognitive skills measures allowing for a more robust analysis of non-cognitive skills and grade variance.

For both the United States and Norway, grade variance is found to be negatively associated with educational outcomes. In the NLSY79, grade variance is negatively associated with educational attainment. In the NRD, grade variance is negatively associated with (1) starting the academic track in upper secondary, (2) upper secondary grade point average, (3) graduating from the academic track in upper secondary and (4) continuing on to higher education. Estimates are robust to controlling for socioeconomic characteristics and school fixed effects in the NLSY79 and school by cohort fixed effects in the NRD. The estimate for grade variance is negative across the grading distribution for both countries and no significant differences are found between boys and girls.

The association between grade variance and educational outcomes remains negative when including measures of cognitive and non-cognitive skills. In the NLSY79, the estimate for grade variance is reduced when adding cognitive skills but remains unchanged when adding non-cognitive skills. In the NRD, adding cognitive and non-cognitive measures do not change results in a systematic way. The CDAP data confirm that grade variance does not seem to be related to non-cognitive skills. While the negative association between grade variance and educational attainment supports the non-cognitive skills hypothesis, all results are robust to adding measures of non-cognitive skills which does not support this hypothesis. My results support the alternative hypothesis that being a generalist rather than a specialist is beneficial for educational attainment.

The paper proceeds as follows. Section 2 discusses why one might expect grade variance

to matter. Section 3 presents the main analysis for the NLSY79 data while section 4 presents the main analysis for the Norwegian register data. Section 5 investigates whether the importance of grade variance depends on the grading distribution, gender and cognitive and non-cognitive skills using all data sources. Section 6 presents the conclusion.

2 Grade variance

Standardized tests, such as the PISA test and the SAT,[1] are designed to be able to determine a student's skills in the specific subject relative to all other students. Grades, however, are a much more subjective measure. Grades are usually decided by the teacher of the subject, are not standardized across classes and schools and can be absolute measures or measured relative to classmates. They are often a combination of knowledge in the subject (cognitive skills) and other skills such as showing up to and participating in class (non-cognitive skills) (Borghans, Duckworth, Heckman, and Ter Weel, 2008; Segal, 2012; Kautz, Heckman, Diris, ter Weel, and Borghans, 2014). In addition, the degree to which cognitive or non-cognitive abilities matter will depend on the subject. Falch, Nyhus, and Strøm (2014), for instance, use math and science grades in school as a proxy for cognitive skills while they use grades in physical education, food and health, arts and crafts and music as a proxy for non-cognitive skills.

On the one hand, grades might reflect non-cognitive skills, such as motivation, perseverance and conscientiousness [2]. Non-cognitive skills have been shown to be meaningful predictors of educational, labor market and behavioral outcomes (Kautz, Heckman, Diris, ter Weel, and Borghans, 2014; Heckman, Stixrud, and Urzua, 2006; Borghans, Duckworth, Heckman, and Ter Weel, 2008; Carneiro, Crawford, and Goodman, 2007; Falch, Nyhus, and Strøm, 2014). Also, non-cognitive abilities have been shown to be more important for the lower part of the skill distribution (Lindqvist and Vestman, 2011). If high grade variance is associated with low non-cognitive skills while low grade variance is associated with high non-cognitive skills, then high grade variance is expected to be associated with low educational attainment, especially at the lower end of the grading distribution. This is the non-cognitive skills hypothesis.

On the other hand, grades might reflect knowledge in the subject, i.e. cognitive skills. High grade variance students have both good and bad skills (specialists) while low grade

[1] The Programme for International Student Assessment (PISA) is a standardized test carried out every three years among a representative sample of 15 year olds, and measures their competency in mathematics, reading and science. Around 510,000 students in a total of 65 countries participated in PISA in 2012 (OECD, 2015). The SAT is a standardized test developed to test students' academic readiness for college. The SAT, along with the ACT, form a large part of the admission decision for many colleges (ACT, 2015; SAT, 2015).

[2] Non-cognitive skills are referred to as soft skills, personality traits, non-cognitive skills, non-cognitive abilities or character and socio-emotional skills, among others. Heckman and Kautz (2013) refer to them as character skills, rather than traits, as they are constant at any age but may change over time. Character skills include "conscientiousness, perseverance (grit), self-control, trust, attentiveness, self-esteem, self-efficacy, resilience to adversity, openness to experience, empathy, humility, tolerance of diverse opinions and the ability to engage productively in society" (Heckman and Kautz, 2013, p. 6).

variance students have more similar skills across subjects (generalists). As higher education allows students to specialize in their preferred field, high variance students might be expected to have a higher educational attainment. This might be especially true for students at the upper end of the grade distribution as these students are more likely to go on to higher education. This is the generalist/specialist hypothesis.

However, it is not clear that being a specialist is always most beneficial. It might be beneficial to be a generalist for some studies or occupations (Lazear, 2004) or it might be beneficial to be a generalist in the long run due to greater adaptability (Hanushek, Woessmann, and Zhang, 2011). Lazear (2004) finds that individuals with balanced skills (jacks-of-all-trades) are more likely to become entrepreneurs. The idea is that rather than having a comparative advantage in a specific skill, entrepreneurs have a comparative advantage in having a span of skills, which is necessary to be successful as an entrepreneur. Being a jack-of-all-trades might be beneficial for the educational outcomes studied in this paper. Higher education is often based on general knowledge suggesting that generalists might be better at higher education. This could particularly be true in the United States where there is a long tradition for a liberal arts education in four-year colleges. The specialist might therefore see the benefit of a short specialized education rather than a long general one. Hanushek, Woessmann, and Zhang (2011) study the impact of vocational versus general education, and find that although individuals with vocational education have an early labor-market advantage due to for instance higher employability, these gains are often offset by reduced adaptability later in life. Being a generalist could be more beneficial for long-run outcomes due to greater adaptability. This is the reversed generalist/specialist hypothesis.

It might also be the case that the association between grade variance and educational attainment differs by gender. A common finding is that while average skill differences between boys and girls tend to be small, the variance of skills is higher for boys than for girls.[3] Although variance across individuals is higher among boys than girls, there is no reason to believe that individual variance is higher for boys than for girls. Even if individual grade variance is higher for boys, it does not necessarily mean that the association between grade variance and education attainment, conditional on grade point average, varies by gender. However, if grade variance to a greater degree reflects being a generalist or specialist for one gender, while it reflects high or low non-cognitive skills for the other gender, results may differ for boys and girls.

Finding a negative association between grade variance and educational attainment, especially at the lower end of the grading distribution, supports the non-cognitive skills hypothesis while finding a positive association, especially at the upper end of the grading distribution, supports the generalist/specialist hypothesis. Also, results could differ by gender if grade variance reflects being a generalist or specialist for one gender, while it

[3]Hedges and Nowell (1995) study six representative large scale surveys with data on mental abilities and find that although average sex difference generally are small, males consistently have larger variance in test scores.

reflects high or low non-cognitive skills for the other gender.

Finally, measures of cognitive and non-cognitive skills are added to the analysis. Grade point average might not be the best measure of cognitive skills. Roth, Becker, Romeyke, Schäfer, Domnick, and Spinath (2015) investigate the relationship between standardized intelligence tests and school grades employing a psychometric meta-analysis and find a population correlation of $\rho = .54$, suggesting that grade point average only proxies as a measure for cognitive skills. Adding improved measures of cognitive skills might therefore strengthen the analysis. Non-cognitive skills are added to the analysis to see whether they explain part of the association between grade variance and educational outcomes. If they do, this suggests that grade variance is capturing a measure of non-cognitive skills and supports the non-cognitive skills hypothesis. If a negative association is found between grade variance and educational outcomes, but results remain unchanged when adding non-cognitive skills, we are left with the reversed generalist/specialist hypothesis that being a generalist rather than a specialist is beneficial for educational attainment. These potential mechanisms are investigated in Section 5.

One concern is that even if we find an association between grade variance and educational attainment, the coefficient for grade standard deviation is picking up a mechanical correlation between grade standard deviation and grade point average due to for instance ceiling effects. By controlling for grade point average, the analysis compares students with the same grade point average, but with different grade variance. However, ceiling effects could affect the association at the lower or upper end of the grading distribution. To investigate whether we are picking up such mechanical effects, the samples are separated into medians and quartiles and separate regressions are run. Finding similar results across all samples removes much of the concern for ceiling effects. Also, in the Norwegian sample, students are bunched at certain values of grade point average where they have exactly the same grade point average but different grade variance. Running a regression for each of these values isolates grade variance from grade point average. Again, finding similar results for all subsamples removes much of the concern for ceiling effects. For more details and results, see Section 5.

3 Grade variance in the United States

In the following, the main results from the National Longitudinal Survey of Youth, 1979 (NLSY79) are presented. The NLSY79 is a longitudinal survey with a nationally representative sample of young Americans and includes high school transcript data, educational attainment and socioeconomic characteristics.

3.1 Institutional setting in the United States

Each state is divided into several school districts, which have jurisdiction over school curricula, budgets and policies for the public schools. State governments set the overall

educational standards and funding for education is a combination of funding from the federal, state and local government. About 10% of students attend private schools (National Center for Education Statistics, 2015) which are free to determine their own curriculum. Compulsory education varies by state, starting between ages five and eight and ending between ages 16 and 18, and may be completed in public schools, private schools or though approved home school programs. Most schools divide their schooling into three levels: elementary school, middle school and high school. "There is no uniform configuration throughout the country in the organization of primary and secondary education. Elementary school begins with kindergarten, but may continue through grades 5, 6, or 8 ... High school typically begins at grade 9 or 10, with middle or junior high schools usually covering the intervening years between elementary school and high school. Students graduate from high school following grade 12". (Stevenson and Nerison-Low, 2002, pp. 15-16) Usually, children are divided into grades by age groups, starting with kindergarten, and then continuing from grades 1 (age 6) to 12 (age 17), where grade 12 is the final year of high school.

A student completing high school will receive a high school diploma, while those students who have not completed high school, or do not meet the requirements for the diploma, have the option of passing a General Education Development (GED) test, a high school equivalency credential. After high school, students may continue on to post-secondary education at colleges or universities. When applying to higher education, the major determinants for admission are grades in college preparatory courses, test scores from the ACT or SAT, and overall grades. Class rank, an application essay or writing samples and letters of recommendation may also be admission criteria (Clinedinst and Hawkins, 2011). Colleges are usually either two-year colleges (community college or junior college) or four-year colleges. Two-year colleges provide academic, vocational and professional education rewarding associate degrees and some students will transfer on to a four-year college. Four-year colleges usually reward a bachelor degree qualifying students for graduate schools where master and doctoral degrees are rewarded.

With this as the institutional background, the analysis uses data on grades received in high school and data on educational attainment, measured as years of completed schooling. A high school degree is equivalent to 12 years of completed schooling while completing a four year college is equivalent to 16 years of completed schooling.

3.2 Data from the National Longitudinal Survey of Youth, 1979

The National Longitudinal Survey of Youth, 1979 (NLSY79) is a survey with a nationally representative sample of 12,686 young Americans between ages 14 and 22 who were first interviewed in 1979. The survey collects information on parental background, schooling decisions, labor market experiences, cognitive and non-cognitive test scores and other behavioral measures on an annual basis. Between 1980 and 1983, transcript information was collected with data on each grade received during high school. See Appendix A for a

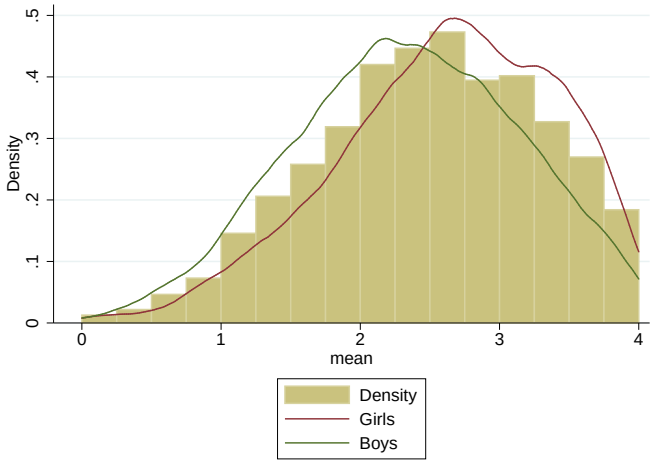
detailed description of the transcript data.

The sample consists of three sub samples: (1) a cross sectional sample of 6,111 respondents from the non-institutionalized segment of the population (2) a supplemental sample of 5,295 Hispanic, Latino, black and economically disadvantaged non-black/non-Hispanic respondents, and (3) a sample of 1,280 respondents enlisted in the military as of September 30, 1978. Following the 1984 interview, most of sample (3) and parts of sample (2) were dropped from the survey. Following Heckman, Stixrud, and Urzua (2006), the main sample with 6,111 respondents is used in the analysis.

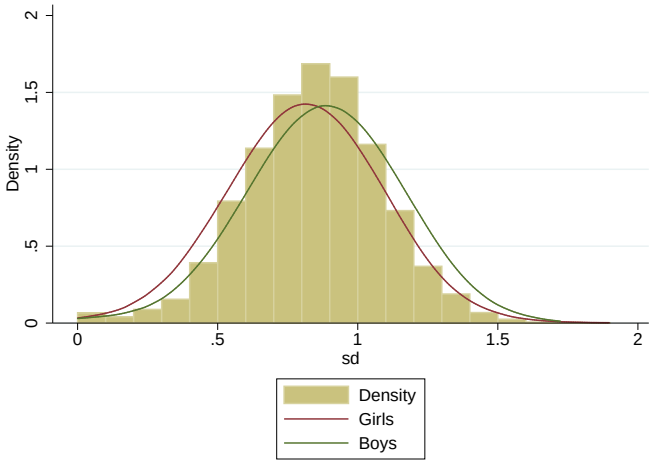
Grade point average (GPA) is measured as the unweighted mean of all grades received in all years of high school (grades 9-12), and is restricted to students with at least 10 valid grades. Grade variance is measured as the standard deviation of an individual's grades (GSD), using the same grades as were used to calculate the individual's grade point average. Descriptive statistics for the transcript data are reported in panel A of Table 1 with the last columns presenting descriptive statistics for girls and boys separately. In the regressions, both GPA and GSD are standardized with mean 0 and standard deviation 1 to facilitate interpretation. The average number of grades is 25.8 with a standard deviation of 6.32. Students either receive a pass/fail grade, or grades A-F, where grade A is coded to value 4, grade B to value 3, grade C to value 2, grade D to value 1. Grade F is a failing grade and is coded to value 0. Figures A1 and A2 in Section A display the distribution of grades and the distribution of number of grades respectively.

Figure 1a displays the distribution of GPA with the red and green lines displaying kernel densities with a bandwidth of 0.15 for girls and boys respectively. Average GPA is higher for girls (2.62) than for boys (2.33) while the spread is slightly higher for boys (standard deviation of GPA is 0.79 for girls and 0.81 for boys). These are both common findings in the literature (Herrnstein and Murray, 2010). Figure 1b displays the distribution of the GSD. Once again, red and green lines displaying kernel densities with a bandwidth of 0.15 for girls and boys respectively. Average GSD is higher for boys (0.88) than for girls (0.81) while the spread in GSD is the same (standard deviation of GSD is 0.24 for both girls and boys).

Figure 1: NLSY79



(a) Distribution of grade point average



(b) Distribution of grade standard deviation

Note: The figure includes 4,389 students from the main sample with 10 or more valid grades and with non-missing educational attainment at age 30. For grade point average, each bin has a width of 0.25, while each bin has a width of 0.1 for grade standard deviation. Lines display kernel densities with bandwidth 0.15 for each variable for girls (red) and boys (green).

The outcome of interest is educational attainment and is measured as years of education at age 30, measured from 1 in 1st grade to 20 in the 8th year of college. Average years of education is 13.5 with a standard deviation of 2.22 (Panel B of Table 1). Educational attainment is similar for boys and girls, while the standard deviation is higher for boys (2.33 for boys and 2.11 for girls). Socioeconomic characteristics include number of siblings, father's highest grade completed, mother's highest grade completed and family income in 1979 as well as a dummy for broken home at age 14, a dummy for living in the south at age 14 and a dummy for living in an urban area at age 14, race and ethnicity dummies. Cohort fixed effects are included in all specifications where cohort corresponds to birth year. The measures of socioeconomic characteristics correspond to those in Heckman, Stixrud, and

Urzua (2006). Descriptive statistics are listed in panel C of Table 1 with last columns of Table 3 presenting descriptive statistics for girls and boys separately.

3.3 Empirical strategy and results

Ideally, we would like to have exogenous variation in grade variance to capture the causal effect of grade variance on educational attainment. However, it is hard to find such variation. Instead, the association between GPA and GSD is estimated using an OLS model controlling for socioeconomic characteristics and including cohort fixed effects. In order to interpret this model as causal, all relevant variables that are correlated with both GSD and educational attainment must be included in the analysis, which is likely not the case. This model therefore expresses the association between GSD and educational attainment, conditional on socioeconomic characteristics and cohort fixed effects.

The outcome variable, y_{it}, is years of education by age 30 for individual i born in year t. GPA_{it} is grade point average and GSD_{it} is grade standard deviation, where each variable is standardized with mean 0 and standard deviation 1. The model includes individual socioeconomic characteristics, $X_t^{'}$, listed in Table 1, and cohort fixed effects, δ_t, in correspondence with Heckman, Stixrud, and Urzua (2006). The error term, ϵ_{it}, is clustered at the cohort level. The model can be expressed as

$$y_{it} = \alpha GPA_{it} + \gamma GSD_{it} + X_{it}^{'}\beta + \delta_t + \epsilon_{it} \tag{1}$$

The variable of interest is γ, which is the conditional correlation of GSD and outcome y, once GPA and other variables are controlled for. If γ is positive, a student with the same GPA but with higher GSD is expected to have more years of education by age 30 whereas a negative γ indicates the opposite.

The results are presented in Table 2 where all columns include cohort fixed effects. The first two columns present a simple OLS regression with GPA as an explanatory variable with and without socioeconomic characteristics. As expected, GPA is positively correlated with educational attainment, with a one standard deviation increase in GPA predicting 1.2 years more of education by age 30. This corresponds to 0.55 of a standard deviation increase in years of education. The estimate remains stable when controlling for socioeconomic characteristics.

In the next columns, the variable of interest, GSD, is added to the model. The coefficient for GSD in columns (3) and (4) tells us how grade standard deviation predicts educational attainment when controlling for GPA. The coefficient for grade standard deviation is -0,242 without controlling for socioeconomic characteristics and -0.238 when controlling for socioeconomic characteristics, indicating that the result is not driven by some sub-sample of students. The coefficient for GPA is only slightly lower when including GSD in the specification.

In the NLSY79, results show that for a given grade point average, students with higher variance complete fewer years of education than students with low grade variance. If GSD

Table 1: NLSY79 - Descriptive statistics

	Total		Boys		Girls	
	mean	(sd)	mean	(sd)	mean	(sd)
A. Transcript data						
Grade point average (GPA)	2.48	(0.81)	2.33	(0.81)	2.62	(0.79)
Grade standard deviation (GSD)	0.84	(0.25)	0.88	(0.24)	0.81	(0.24)
Number of grades	25.8	(6.32)	25.6	(6.44)	26.0	(6.20)
B. Outcome variable						
Years of education	13.5	(2.22)	13.5	(2.33)	13.6	(2.11)
C. Socioeconomic characteristics						
Girl	0.51	(0.50)	0	(0)	1	(0)
Black	0.11	(0.31)	0.11	(0.31)	0.11	(0.31)
Hispanic	0.061	(0.24)	0.061	(0.24)	0.061	(0.24)
Living in south	0.30	(0.46)	0.29	(0.45)	0.32	(0.47)
Living in urban area	0.76	(0.43)	0.76	(0.43)	0.76	(0.43)
Broken home	0.22	(0.41)	0.21	(0.41)	0.22	(0.41)
Number of siblings	3.20	(2.14)	3.14	(2.13)	3.25	(2.15)
Month of birth	6.45	(3.38)	6.49	(3.41)	6.41	(3.34)
Family income 1979 (thousands)	17.0	(15.1)	17.7	(15.3)	16.4	(15.0)
Mother: Years of education	11.3	(3.47)	11.3	(3.62)	11.4	(3.33)
Father: Years of education	11.2	(4.64)	11.3	(4.69)	11.1	(4.59)
D. Cognitive skills						
Arithmetic reasoning (ASVAB 1)	18.2	(7.19)	19.2	(7.34)	17.3	(6.92)
Word knowledge (ASVAB 2)	26.4	(7.12)	26.3	(7.35)	26.5	(6.90)
Paragraph comprehension (ASVAB 3)	11.2	(3.17)	10.8	(3.34)	11.5	(2.97)
Mathematical knowledge (ASVAB 4)	46.6	(15.2)	42.8	(14.8)	50.3	(14.7)
Coding speed (ASVAB 5)	14.1	(6.31)	14.4	(6.50)	13.8	(6.10)
Cognitive	0	(1.00)	-0.042	(1.05)	0.041	(0.95)
E. Non-cognitive skills						
Rotter locus of control scale	7.56	(2.38)	7.62	(2.36)	7.50	(2.39)
Rosenberg self-esteem scale	22.7	(4.05)	22.9	(3.96)	22.5	(4.12)
Non-cognitive	0	(1.00)	0.046	(0.98)	-0.045	(1.02)

Note: N=4,389 for the whole sample, with 2,234 girls and 2,155 boys. N=4,243 for the cognitive measure and N=4,225 for the non-cognitive measure. N=4,136 when combining the cognitive and non-cognitive measures.

Table 2: NLSY79 - Years of education by age 30

	(1)	(2)	(3)	(4)
Grade Point Average	1.223***	1.096***	1.092***	0.972***
	(0.029)	(0.026)	(0.032)	(0.028)
Grade Standard Deviation			-0.242***	-0.238***
			(0.033)	(0.034)
Socioeconomic Characteristics	No	Yes	No	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.304	0.391	0.312	0.399
N	4,389	4,389	4,389	4,389

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the cohort level.

increases by one standard deviation, educational attainment is reduced by 1/4 of a year. This corresponds to 0.11 of a standard deviation decrease in years of education.

4 Grade variance in Norway

To provide a similar and comparable investigation of Norway, I use Norwegian Register Data (NRD). Comparing results from the NRD to those from the NLSY79 indicates whether the results are country and context specific or more general. For instance, upper secondary and higher education in Norway has a high degree of tracking, which is not the case for the United States. According to the specialist/generalist hypothesis, high grade variance might therefore be associated with high educational attainment in Norway and the opposed to the United States.

Using the NRD has clear benefits. Firstly, the data cover the entire student cohort for three years, a sample of over 150,000 students. Secondly, the data include school identifiers so that school by cohort fixed effects can be added to the analysis. Thirdly, admission into upper secondary education and higher education is centralized and almost entirely based on GPA. It is therefore less likely that important variables are omitted from the analysis when including GPA and socioeconomic characteristics along with GSD as the only measures determining educational attainment. Lastly, grading is monitored by the central government which reduces concerns of measurement error.

4.1 Institutional setting in Norway

There are clear institutional differences between Norway and the United States. In Norway, municipalities (428) are responsible for primary and lower secondary education, while counties (19) are responsible for the upper secondary education. Compulsory education consists of primary education (grades 1-7) and lower secondary education (grades 8-10), and ends the year the student turns 16 years of age, and entrance into primary and lower secondary education is determined by catchment areas. There is no possibility to fail a class in primary or in lower secondary education during the empirical period, implying

that all students finish compulsory education on time.[4] There is no tracking, a common national curriculum for all students and very few private schools, with only 3.5 % of students attending a private elementary or lower secondary school in 2015 (The Norwegian Directorate for Education and Training, 2015).

Children do not receive grades in primary education.[5] In lower secondary education, students receive grades from their teachers every semester, primarily based on their performance in the subject. These grades have no consequences for the students prior to grade 10. Grades received in the last semester of grade 10, along with 2-3 externally graded oral or written exams, are used to determine acceptance to upper secondary education. Students are only tested in theoretical subjects on the exams, and the subject to be tested is decided by a draw. The written exams are the same nationally for all students taking the specific subject, while the oral exams are organized locally. The externally-graded grades are averaged with the teacher-graded grades in the corresponding subjects. The unweighted grade point average of the resulting grades is used to determine acceptance into upper secondary education.

Students may choose from 3 study tracks qualifying for higher education, and 12 vocational study tracks. When applying for upper secondary education, students rank their preferred study tracks and schools within study tracks. All students have been guaranteed admission to upper secondary education since 1994, but whereas acceptance to one of their three ranked choices is guaranteed, the grade point average determines which school and study program the student is accepted to. How important grades are for entering the school or study program of their choice will vary from county to county as counties are free to determine how acceptance into upper secondary education is organized (Haraldsvik, 2003).

In upper secondary education, academic tracks have a duration of 3 years while vocational tracks typically last for 4 years, including 2 years of apprenticeship training. Subject requirements differ depending on the study program and there are both mandatory and elective subjects. If students from vocation tracks want to continue on to higher education, they can attend a year of supplementary studies qualifying for higher education.

The application system to higher education is centralized for the entire country and is solely based on grade points.[6] There are two application categories. In the first category, grade points are calculated using grade point average and any science or advanced placement credits if applicable. In the second category, grade points include any attempts at grade improvements and adds credits for e.g. age, military service, years of study in higher education. Students automatically apply in both categories, but most students are

[4] In very few cases, students do not start primary education at the expected age, which implies that they finish lower secondary education at different age. If a child is not considered to be mature enough, the parents together with the school and psychologists can postpone enrollment one year. In addition, some older students return to improve their grades, and immigrants are often over-aged at graduation.

[5] Students in the highest grades of elementary education will in some cases receive grades as preparation for lower secondary education. The grades have no direct consequences for the students.

[6] There are only some exceptions, such as music and architecture where admissions are determined by an entrance exam as well.

accepted in the first category. In both cases, grade point average is the major determinant of acceptance into higher education.

The major difference between Norway and the United States is that Norway has a much more centralized educational system. There is a national curriculum, in contrast to the United States where states and school districts have more influence. Although some students do attend private schools in Norway, they are highly regulated. In Norway, there is a centralized system for applying to higher education whereas each institution decides their admission criteria in the United States. Due to the centralized system, grading in Norway is monitored by the central government which reduces concerns of measurement error in the analysis.

In the following, grades from lower secondary school are used in the analysis. Educational outcomes are related to whether the student starts academic or vocational track in upper secondary education, grades in upper secondary education, whether the student completes upper secondary education and whether the student continues on to higher education.

4.2 Norwegian register data

Using register data, provided by Statistics Norway for all individuals leaving lower secondary education during 2002-2004, allows for the combination of detailed information on individual's background and education, including grades, measures of educational attainment and socioeconomic characteristics. The sample is restricted to students with at least 10 valid teacher-assessed grades and only includes students graduating from lower secondary education at age 16.[7] Also, students must have non-missing information on the lower secondary school they attended. The data reduction is presented in Table B1.

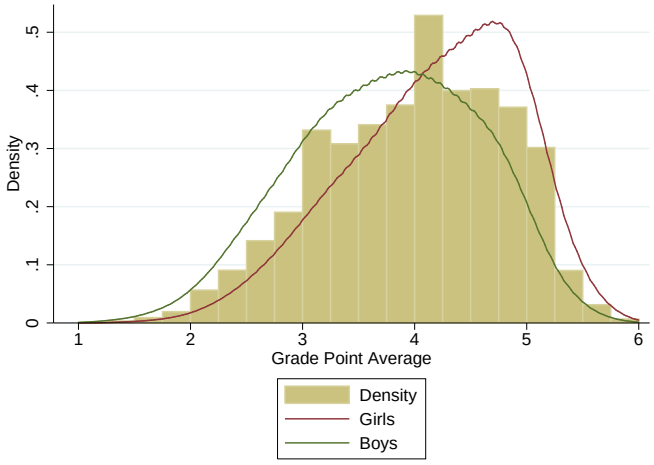
Grade point average (GPA) in the NRD is measured as the unweighted mean of all 13 teacher-assessed grades received when leaving lower secondary education. The subjects are written and oral Norwegian, written and oral English, mathematics, natural science, social science, religion, home economics, music and arts, physical education and crafts. Grade variance is measured as the standard deviation of an individual's grades (GSD), using the same grades as were used to calculate the individual's grade point average. Descriptive statistics are presented in panel A of Table 3. In the regressions, both variables are standardized with mean 0 and standard deviation 1 to facilitate interpretation. About 90 % of students in the sample have 13 valid grades. Figure B1 in Section B displays the distribution of grades from one (the lowest) to six (the highest). The most common grade is four (34%), while the least common grade is one (0.86%).

Figures 2a and 2b are equivalent to Figures 1a and 1b of Section 3.2. The distributions are remarkably similar to the NLSY79: The distributions of GPA are skewed to the right

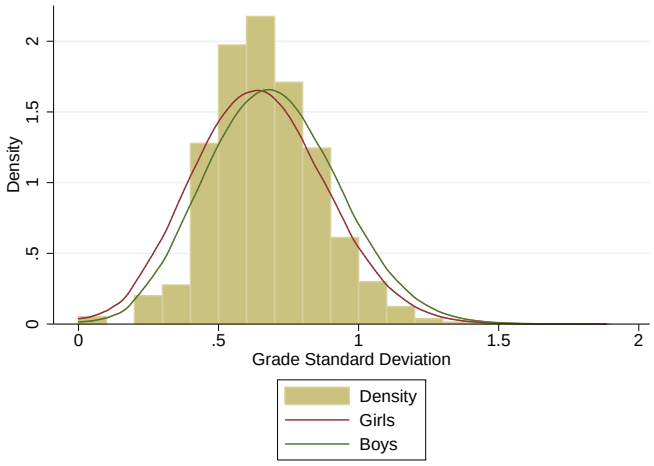
[7] In some cases, students do not start primary education at the expected age, which implies that they finish lower secondary education at a higher age. If a child is not considered to be mature enough, the parents together with the school and psychologists can postpone enrollment one year. In addition, some older students return to improve their grades, and immigrants are often over-aged at graduation.

while the distributions of GSD are approximately normally distributed. Also, the gender differences are identical to the NLSY79. GPA is higher for girls (4.18) than for boys (3.77) and GSD is higher for boys (0.69) than for girl (0.64). This suggests that the measures of both GPA and GSD are comparable in general and for both genders across countries.

Figure 2: NRD



(a) Distribution of grade point average



(b) Distribution of grade standard deviation

Note: For grade point average, each bin has a width of 0.25, while each bin has a width of 0.1 for grade standard deviation. Lines display kernel densities with bandwidth 0.15 for each variable for girls (red) and boys (green).

In the NLSY79, the main outcome variable was years of education at age 30. The analysis in the NRD uses cohorts leaving lower secondary school in Norway in 2002-2004, as 2002 is the first year grade data became available. In the last data point available, 2011, these students were still too young to have completed all years of higher education. Therefore, measures for educational attainment in the NRD are short-run measures and are measured as (1) Started academic track (Started ACA), (2) vocational track gradu-

ate (VOC grad) (3) academic track graduate (ACA grad) (4) grade point average upper secondary education (GPA UPE) and (5) started higher education (Started HE). Started academic track is an indicator variable equal to one if the student started one of the three academic study tracks in the first year of upper secondary education. 97% of students go on to upper secondary education in the fall after completing lower secondary education, with 46% starting an academic track and 51% starting vocational tracks. Vocational track graduate and academic track graduate are indicator variables equal to one if the student starts vocational or academic upper secondary education and graduates within five years. Students have a legal right to five years of upper secondary education and this is the standard measure for upper secondary education completion used by the authorities. 70% of students graduate from upper secondary education within five years. Grade point average upper secondary education (GPA USE) is measured as the unweighted mean of all teacher-assessed grades on the upper secondary education transcript, standardized with mean 0 and standard deviation 1. The measure only includes students who complete the academic track and have at least 10 valid grades. Students who transfer from the vocational to the academic track are also included. GPA USE has a mean of 4.15 and a standard deviation of 0.68. The last measure, started higher education, is an indicator variable equal to one if a student has started, but not necessarily completed, a higher education program before 2012. 53% of the sample start higher education. Descriptive statistics are presented in panel B of Table 3.

Socioeconomic characteristics in the NRD are quite similar to the NLSY79. They include gender, birth month, immigration status,[8] parental employment status[9] and parental education.[10] Variables are measured the year the student turns 16. Descriptive statistics are presented in panel C of Table 3. The last columns of Table 3 present descriptive statistics for girls and boys separately. Boys are less likely to start the academic track, have lower GPA and higher GSD in upper secondary education, are less likely to complete upper secondary education and less likely to start higher education.

[8]Immigration status is divided into two categories, where the first indicates that you are a first generation immigrant born abroad with parents born abroad and the second indicates that you are a second-generation immigrant, born in Norway but with both parents born abroad.

[9]Parental employment status is an indicator for whether only the mother, only the father or both parents are working, where no parents working is the reference category.

[10]Parental education as measured as the highest completed education by one of the parents, with categories including having completed upper secondary education, a Bachelor's degree, a Master's degree or PhD and having an unknown education, with less than upper secondary education being the reference category.

Table 3: NRD - Descriptive statistics

	Total		Boys		Girls	
	mean	(sd)	mean	(sd)	mean	(sd)
A. Transcript data						
Grade Point Average (GPA)	3.97	(0.82)	3.77	(0.82)	4.18	(0.77)
Grade Standard Deviation (GSD)	0.67	(0.19)	0.69	(0.19)	0.64	(0.19)
Number of grades	12.87	(0.42)	12.84	(0.47)	12.90	(0.37)
B. Outcome Variables						
Started academic track	0.46	(0.50)	0.42	(0.49)	0.50	(0.50)
Vocational track graduate	0.60	(0.49)	0.56	(0.50)	0.64	(0.48)
Academic track graduate	0.85	(0.36)	0.81	(0.39)	0.88	(0.33)
GPA upper secondary education	4.15	(0.68)	4.06	(0.69)	4.21	(0.67)
Started higher education						
- complete sample	0.53	(0.50)	0.43	(0.49)	0.63	(0.48)
- academic track	0.88	(0.33)	0.88	(0.32)	0.88	(0.33)
C. Socioeconomic characteristics						
Girl	0.49	(0.50)				
Birth month	6.41	(3.36)	6.39	(3.35)	6.44	(3.37)
First generation immigrant	0.034	(0.18)	0.034	(0.18)	0.034	(0.18)
Second generation immigrant	0.020	(0.14)	0.020	(0.14)	0.021	(0.14)
Parental education: Upper secondary	0.47	(0.50)	0.47	(0.50)	0.47	(0.50)
Parental education: Bachelor	0.29	(0.45)	0.29	(0.45)	0.29	(0.45)
Parental education: Master +	0.10	(0.30)	0.10	(0.30)	0.10	(0.30)
Parental education: Unknown	0.042	(0.20)	0.042	(0.20)	0.042	(0.20)
Only mother working	0.13	(0.34)	0.13	(0.33)	0.13	(0.34)
Only father working	0.12	(0.33)	0.13	(0.33)	0.12	(0.33)
Both parents working	0.68	(0.47)	0.68	(0.47)	0.68	(0.47)
D. Cognitive and non-cognitive skills						
Cognitive skills	3.68	(1.06)	3.57	(1.08)	3.81	(1.04)
Non-cognitive skills	4.26	(0.72)	4.09	(0.73)	4.44	(0.67)

Note: N=158,308, with 80,701 boys and 77,607 girls. For Grade point average upper secondary education and grade standard deviation upper secondary, N=84,010 with 33,334 boys and 50,676 girls.

4.3 Empirical strategy and results

For the NRD, the estimated model is equivalent to the one estimated using the NLSY79 data, except that school by cohort fixed effects, $\delta_t \times \theta_s$, are added. y_{ist} is the outcome for student i from school s in year t. GPA_{ist} is grade point average and GSD_{ist} is grade standard deviation from lower secondary education, where each variable is standardized with mean 0 and standard deviation 1. X_{ist} is a vector of socioeconomic characteristics including gender, immigrant status, parental education, parental employment status and birth month. Socioeconomic characteristics are listed in Table 3. The error term ϵ_{ist} is clustered at the school level. The model can be expressed as

$$y_{ist} = \alpha GPA_{ist} + \gamma GSD_{ist} + X'_{ist}\beta + \delta_t \times \theta_s + \epsilon_{ist} \quad (2)$$

Table 4 reports the results where the outcome is the indicator variable for whether the student has started higher education. The table is equivalent to Table 2 in Section 3.3, with the exception that school by cohort fixed effects are added to the last column. As with the NLSY79, GPA is as expected positively correlated with the educational outcome. Increasing GPA by one standard deviation increases the likelihood that one starts higher education by 30%, which is equivalent to 0.6 of a standard deviation and is similar to the finding for NLSY79.

GSD is added in Column (3) and is negatively correlated with starting higher education. A one standard deviation increase in GSD decreases the likelihood that one starts higher education by 3.2%. This is equivalent to 0.06 of a standard deviation increase in the likelihood of starting higher education. This is approximately half of the GSD estimate found for years of education in the NLSY79. The results remain remarkably stable when adding socioeconomic characteristics (Column (4)) and school by cohort fixed effects (Column (5)), indicating that neither student background nor school characteristics are driving the results.

Table 4: NRD - Started higher education

	(1)	(2)	(3)	(4)	(5)
GPA	0.306***	0.270***	0.292***	0.257***	0.266***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
GSD			-0.032***	-0.030***	-0.027***
			(0.001)	(0.001)	(0.001)
Soc. Char	No	Yes	No	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	No
CohortxSchool FE	No	No	No	No	Yes
R-squared	0.375	0.402	0.380	0.405	0.397
N	158,308	158,308	158,308	158,308	158,308
Number of groups					3,397

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Started higher education is an indicator variable equal to 1 if the student has started higher education before 2012. Standard errors are clustered at the school level.

Table 5 displays results for the outcome variables described in Section 4.2. The last column is equivalent to Column (5) of Table 4, expect that only students graduating from the academic track are included. All estimations include socioeconomic characteristics and school by cohort fixed effects. Estimates show that GSD is negatively associated with starting the academic track, graduating from the academic track, upper secondary grade point average and starting higher education. The estimate for graduating from upper secondary for students starting the vocational track is small and insignificant. The estimate for GSD in Table 4 seems to be the combined result of students with higher GSD (1) having a higher probability of starting vocational track, where one is less likely to go on to higher education and (2) being less likely to graduate from the academic track and (3) receiving lower grades in the academic track.

Table 5: Main Results - NRD

	Started ACA	VOC grad	ACA grad	GPA USE	Started HE
GPA	0.244***	0.270***	0.210***	1.012***	0.137***
	(0.002)	(0.002)	(0.003)	(0.006)	(0.003)
GSD	-0.018***	-0.0003	-0.013***	-0.012***	-0.004**
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)
Soc. Char	Yes	Yes	Yes	Yes	Yes
CohortxSchool FE	Yes	Yes	Yes	Yes	Yes
R-squared	0.301	0.242	0.223	0.542	0.111
N	158,308	80,725	72,839	83,740	83,740
Number of groups	3,397	3,306	3,194	3,208	3,208

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Academic is an indicator variable for whether the student goes on to academic track. VOC grad includes all students who start vocational track and complete upper secondary within 5 years. ACA grad includes all students who start academic track and complete upper secondary within 5 years. GPA USE is the GPA from upper secondary education for students who have graduated from the academic track of upper secondary school. This includes students who have transferred from the vocational track during upper secondary school. Started HE is an indicator variable for whether the student has started higher education before 2012 and includes the same sample as GPA USE.

Both the results from Norway and the United States show a negative association between grade variance and educational attainment when controlling for GPA. These findings do not support the hypothesis that being a specialist in compulsory education is beneficial for further education. However, it is still an open question whether the relationship between grade variance and educational attainment depends on the grading distribution, gender and cognitive and non-cognitive skills.

5 Grading distribution, gender and skills

How does the relationship between grade variance and educational attainment depends on the grading distribution, gender and cognitive and non-cognitive skills? In answering this question, all analyses below are based on the regression in column (4) of Table 2 for the NLSY79 data and column (5) of Table 4 for the NRD.

Grading distribution

If high grade variance individuals are specialists, grade variance is expected to be positively associated with educational attainment, particularly in the upper end of the grade distribution. If high grade variance individuals are individuals with low non-cognitive skills, grade variance is expected to be negatively associated with educational attainment, particularly in the lower end of the grade distribution.

The following investigates whether the direction or strength of the relationship depends on where the student is located in the grading distribution. Regression results reported in Tables 2 and 5 might be masking such differences. To investigate this hypothesis in the

NLSY79 data, separate regressions are run for observations above and below the median grade point average, and then separately for each quartile of grade point average. The results are presented in Table 6. The first column shows results for observations below the median grade point average, while the second column shows results for observations above. Both coefficients are negative and significant, but the coefficient is much more negative for the sample above the median. The same pattern emerges when the regression is run for each quartile, however results are no longer significant as the standard errors increase due to fewer observations.

Table 6: NYLS79: Years of education by age 30 - median and quartiles

	Below med.	Above med.	Q1	Q2	Q3	Q4
GPA	0.881***	1.222***	0.937***	1.298***	1.035*	1.017
	(0.042)	(0.161)	(0.128)	(0.232)	(0.379)	(0.455)
GSD	-0.095*	-0.245*	-0.130	-0.042	-0.206	-0.284
	(0.032)	(0.088)	(0.056)	(0.031)	(0.114)	(0.189)
Soc. Char	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.230	0.290	0.202	0.138	0.147	0.259
N	2,200	2,189	1,101	1,099	1,098	1,091

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the cohort level.

Table 7 displays the same regressions for the NRD. Once again, the coefficient is negative across all quartiles, and for the NRD, the coefficient is also strongly significant across all specifications. In the NRD, however, it seems to be that the strongest relationship between GSD and educational attainment is at the middle of the grading distribution. The coefficient is -0.028 and -0.20 in the middle quartiles, while the coefficient is -0.015 in the lowest quartile and -0.012 in the highest quartile.

Table 7: NRD: Started higher education - median and quartiles

	Below med.	Above med.	Q1	Q2	Q3	Q4
GPA	0.202***	0.178***	0.069***	0.391***	0.315***	0.098***
	(0.003)	(0.004)	(0.003)	(0.010)	(0.012)	(0.004)
GSD	-0.031***	-0.019***	-0.015***	-0.028***	-0.020***	-0.012***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)
Soc. Char	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	0.174	0.111	0.046	0.112	0.075	0.048
N	84,085	74,223	41,309	42,776	37,138	37,085

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the school level.

As grades in Norway can only take on integer values from one to six, students are bunched at certain values of GPA. When calculating the grade point average, receiving grades two and four is equivalent to receiving two three's which means that although

students have exactly the same GPA, they can have different GSD. This feature not only makes it possible to investigate whether there are heterogeneous results across the grading distribution, it also makes it possible to remove any concern that the coefficient for grade standard deviation is picking up a mechanical correlation between grade standard deviation and grade point average due to for instance ceiling effects.

The analysis is restricted to values where there are at least 1000 students, leaving 38 unique GPA values. Figure 3a displays the mean, minimum and maximum value of GSD for each value of the 38 GPA values. There is a spread in GSD for each value of GPA, which is the variation used to identify how GSD is associated with educational attainment. A separate regression is run at each of these values, and results are reported in Figure 3b. The point estimates are always negative. Confidence intervals show that estimates are lower and significantly different from zero at the middle of the grading distribution, while they are typically not significantly different from zero at the lower and higher end of the grading distribution. This corresponds to the results found in Table 7. The results indicate a negative association between GSD and GPA across the grading distribution, and that this is not solely due to a mechanical correlation between the two variables.

For both the United States and Norway there is no evidence of the direction of the estimates changing across the grading distribution. All point estimates are negative and are significantly lower than zero in most cases. There is also no evidence that the relationship is stronger at the lower part of the grading distribution. In the non-cognitive skills hypothesis and the specialist/generalist hypothesis, grade variance is thought to be particularly important at the lower and upper end of the grading distribution respectively. There is no support for either in the data.

Gender

Does the relationship between GSD and educational attainment depend on gender? The results could differ by gender if for instance grade variance reflects being a generalist or specialist for one gender, while it reflects high or low non-cognitive skills for the other gender.

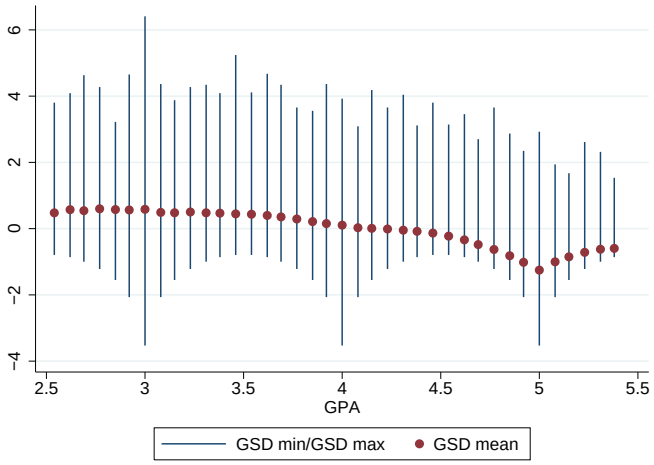
For both the NYLS79 and the NRD, the main estimation is run separately for boys and girls. The results are reported in Table 8. The estimates for GSD are not statistically different between genders in either the NYLS79 (columns (1) and (2)) or in the NRD (columns (3) and (4)). The negative association between GSD and educational attainment is the same direction and magnitude for both genders in the United States and Norway.[11]

These estimations show that the main results are not masking differences across boys and girls. Some might believe that high grade variance reflect low non-cognitive skills for boys while it reflects being a specialist for girls. There is no evidence to support this

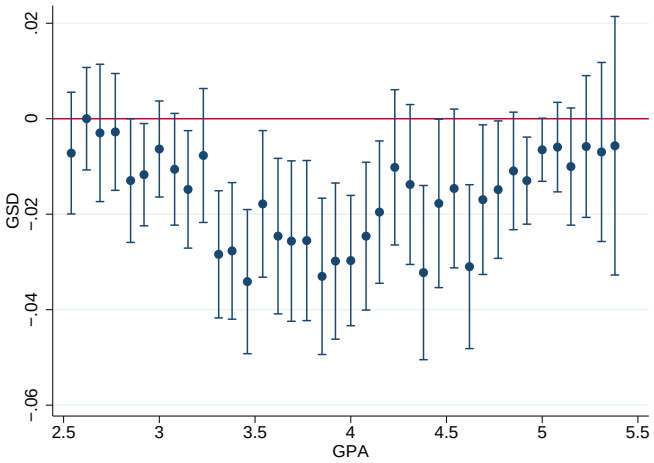
[11] Another way to investigate whether results differ by gender is to see how the coefficient for gender in the regressions that includes socioeconomic characteristics changes when GSD is added to the estimation. For both the NYLS79 and the NRD, the coefficient for female stays the same when adding GSD to the regression. The estimate changes from -0.35 to -0.32 in the NYLS79 and from 0.066 to 0.065 in the NRD.

Figure 3: NRD: Started higher education - grading distribution

(a) GSD - Descriptive figure



(b) GSD - Regression results



Note: GSD is standardized for the entire sample with mean 0 and standard deviation 1. GPA corresponds to the 38 values of grade point average where there are at least 1000 observations. Figure 3a: Dots indicate the mean value while the bars indicate the minimum and maximum vales of GSD for each regression. Figure 3b: Regressions include socioeconomic characteristics and cohort fixed effects. Dots indicate the coefficient for each regression while the bars indicate the 95% confidence interval.

Table 8: NYLS79 and NRD: Results by gender

	NYLS79		NRD	
	Girls	Boys	Girls	Boys
GPA	0.797***	1.131***	0.278***	0.259***
	(0.040)	(0.035)	(0.002)	(0.002)
GSD	-0.259***	-0.234***	-0.023***	-0.028***
	(0.034)	(0.040)	(0.002)	(0.002)
Soc. Char	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	-	-
CohortxSchool FE	-	-	Yes	Yes
R-squared	0.345	0.453	0.377	0.369
N	2,234	2,155	77,605	80,701
Number of groups	-	-	3,287	3,287

* $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$

Note: Standard errors are clustered at the cohort level for the NYLS79 and school level for the NRD.

theory as the estimates are negative for both genders. Also, there is no evidence that grade variance is more important for one gender as the estimates are not statistically different.

Cognitive and non-cognitive skills

How are results affected by including measures of cognitive and non-cognitive skills to the analysis? If grade point average does not perfectly capture cognitive skills (Roth, Becker, Romeyke, Schäfer, Domnick, and Spinath, 2015) then adding improved measures of cognitive skills might reduce a potential bias in the estimate of GSD. Non-cognitive skills are added to see whether they explain part of the association between grade variance and educational outcomes. If they do, this suggests that grade variance is capturing a measure of non-cognitive skills and supports the non-cognitive skills hypothesis.

The analysis is conducted using all three data sources. In the NLSY79 data, measures of cognitive and non-cognitive skills previously used by Heckman, Stixrud, and Urzua (2006) are added to the analysis. In the NRD, measures of a student's skills in cognitive and non-cognitive subjects, based on a subset of subjects, are added to the analysis. Finally, data from the Development in Adolescence Project (CDAP) are used to investigate how non-cognitive skills relate to GSD when conditioning on GPA.

The measure for cognitive skills in the NLSY79 is a composite score of five measures from the Armed Services Vocational Aptitude Battery (ASVAB),[12] which includes scores for arithmetic reasoning, word knowledge, paragraph comprehension, mathematical knowledge and coding speed. Descriptive statistics are reported in panel D of Table 1. For each measure, the scores are standardized with mean 0 and standard deviation 1, and the sum of these five scores is then again standardized with mean 0 and standard deviation 1.

[12]The Armed Services Vocational Aptitude Battery (ASVAB) is a battery of tests administered to applicants to the United States military to determine their qualifications and job assignment. The Armed Forces Qualifying Test AFQT is comprised of test results from the batteries Arithmetic Reasoning, Math Knowledge, Word Knowledge and Paragraph Comprehension (ASVAB, 2015).

Table 9: NYLS79: Conditional correlations

	(1) GSD	(2) GSD	(3) GSD	(4) GSD
GPA	-0.521***	-0.523***	-0.464***	-0.462***
	(0.014)	(0.014)	(0.017)	(0.017)
Non-cognitive		-0.009		0.011
		(0.014)		(0.014)
Cognitive			-0.127***	-0.136***
			(0.019)	(0.020)
Soc. Char	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.302	0.309	0.315	0.318
N	4,389	4,226	4,243	4,136

Note: Standard errors are clustered at the cohort level.

The measure for non-cognitive skills in the NLSY79 is a combination of the Rotter Locus of Control Scale (Rotter, 1966), and the Rosenberg Self-Esteem Scale (Rosenberg, 1965). The Rotter Locus of Control Scale is designed to measure the extent to which individuals believe they have control over their lives through self-motivation or self-determination (internal control) as opposed to the extent that the environment (chance, fate, luck) controls their lives (see Table A2). The Rosenberg Self-Esteem Scale describes ones degree of approval or disapproval toward oneself. (see Table A3). Descriptive statistics are reported in panel E of Table 1 above. Both scores are standardized with mean 0 and standard deviation 1, and the sum of these two scores is then again standardized with mean 0 and standard deviation 1.

Table 9 displays the conditional correlation between GPA and GSD when including cognitive and non-cognitive measures to the NLSY79 data. Column (1) is the conditional correlation between GPA and GSD when including school fixed effects and socioeconomic characteristics. Column (2) adds the measure of non-cognitive skills, column (3) adds the measure of cognitive skills and column (4) adds both measures. We see that the measure for non-cognitive skills is not significant while the measure for cognitive skills is negatively associated with GSD, conditional on GPA. Importantly, adding non-cognitive skills does not change the conditional correlation between GPA and GSD.

In Table 10, cognitive and non-cognitive measures are added to the main analysis. Descriptive statistics are presented in panel D of Table 3. The estimate for non-cognitive skills, as shown in column (2) is significant and positive, as expected, with a one standard deviation increase in non-cognitive skills predicting an increase in educational attainment by 0.26 of a year. However, the estimates for GPA and GSD are unchanged, suggesting that the measure of non-cognitive skills does not explain why GSD is negatively associated with educational attainment. The measure for cognitive skills, as shown in column (3), is significantly and positively associated with educational attainment and reduces both the estimate for GPA and GSD. A one standard deviation increase in cognitive skills predicts

Table 10: NYLS79: Years of education age 30

	(1)	(2)	(3)	(4)
Grade Point Average	0.972***	0.913***	0.604***	0.598***
	(0.028)	(0.033)	(0.030)	(0.031)
Grade Standard Deviation	-0.238***	-0.250***	-0.176**	-0.182**
	(0.034)	(0.036)	(0.036)	(0.034)
Non-cognitive		0.263***		0.143**
		(0.047)		(0.039)
Cognitive			0.800***	0.752***
			(0.044)	(0.041)
Socioeconomic Characteristics	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
R-squared	0.399	0.415	0.461	0.463
N	4,389	4,226	4,243	4,136

* $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$

Note: Standard errors are clustered at the cohort level.

an increase in educational attainment by 0.8 of a year. Column (4) includes both measures, with estimates for GPA and GSD remaining stable from column (3) to column (4). The results correspond to those found for the conditional correlations. The main inference from these estimates is that there is no evidence that the relationship between GPA and GSD or the relationship between GSD and educational attainment can be explained by non-cognitive skills in the NLSY79 data. Also, the estimate for GSD remains negative and statistically significant in all specifications.

In the Norwegian data, measures of a student's skills in cognitive and non-cognitive subjects are added to the analysis. Falch, Nyhus, and Strøm (2014), using the same grade data from Norway as this paper, use the average grade in math and science as a proxy for cognitive skills and the average grade in physical education, food and health, arts and crafts and music as a proxy for non-cognitive skills. These same measures are standardized with mean 0 and standard deviation 1 and added to the analysis to investigate how cognitive and non-cognitive skills relate to GSD in the Norwegian data. Note that these measures are sub-samples of the grades used to calculate GPA and GSD. They are imperfect measures that do not add any new information, but rather take out some of the variation. This makes the results hard to interpret.

Table 11, comparable to Table 9, displays the conditional correlation between GPA and GSD when including these cognitive and non-cognitive measures. The non-cognitive measure is positively associated with GSD while the cognitive measure is negatively associated with GSD. For a given GPA, students with good grades in non-cognitive subjects have higher GSD, while students with good grades in cognitive subjects have lower GSD. The conditional correlation between GSD and GPA is greatly affected by the inclusion of measures of non-cognitive and cognitive skills. This is not surprising as these variables are subsets of grades used to calculate GSD and GPA. However, it is interesting to note that

Table 11: NRD: Conditional correlations

	GSD	GSD	GSD	GSD
GPA	-0.431***	-1.010***	0.054***	-0.586***
	(0.005)	(0.010)	(0.009)	(0.015)
Non-cognitive		0.659***		0.586***
		(0.010)		(0.010)
Cognitive			-0.524***	-0.389***
			(0.009)	(0.009)
Socioeconomic Characteristics	Yes	Yes	Yes	Yes
CohortxSchool FE	Yes	Yes	Yes	Yes
R-squared	0.212	0.305	0.260	0.331
N	158,308	158,308	158,289	158,289
Number of groups	3,397	3,397	3,397	3,397

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the school level.

the different subsets do, in fact, seem to measure something different, but whether this is cognitive and non-cognitive skills is harder to determine.

As the conditional correlations are differentially affected by including measures of cognitive and non-cognitive skills, it is reasonable to assume that this will also be the case when including these measures to the estimations in Table 5. Tables B2 - B6 in Appendix B report the results and this is indeed the case. However, the results are hard to interpret as the measures of cognitive and non-cognitive skills are so closely related to GPA and GSD.

The results from the NLSY79 data show that, if anything, grade variance is associated with cognitive skills rather than non-cognitive skills, while the results from the NRD show no clear pattern. However, both measures of non-cognitive skills are quite simple and do not necessarily include the non-cognitive skills one would associate with low grade variance. To explore this further, data from the Character Development in Adolescence Project (CDAP), provided by Angela Duckworth, are used to investigate the non-cognitive skills in greater detail. The data include grades and a rich set of non-cognitive skills allowing me to investigate how non-cognitive skills relate to GSD when conditioning on GPA (see Section C1 for a description of the data). Non-cognitive skills are either self-reported by the student or reported by the student's teachers. The self-reported measure (Non-cognitive: SR) is a joint measure for the non-cognitive skills (1) delay discounting, (2) grit, (3) self-control: work, (4) self-control: interpersonal, (5) gratitude, (6) actively open-minded thinking, (7) prosocial purpose and (8) internal locus of control. The teacher-reported measure (Non-cognitive: TR) is a joint measure for the non-cognitive skills (1) grit, (2) self-control: work, (3) self-control: interpersonal, (4) gratitude, (5) actively open-minded thinking and (6) prosocial purpose. The results are displayed in Table 12. Column (1) displays the conditional correlation between GPA and GSD which is negative and significant. Column (2) adds the self-reported non-cognitive measure, column (3) adds the

Table 12: CDAP: Conditional correlations

	(1) GSD	(2) GSD	(3) GSD	(4) GSD
GPA	-0.343***	-0.350***	-0.340**	-0.365*
	(0.040)	(0.050)	(0.081)	(0.092)
Non-cognitive: SR		-0.012		-0.013
		(0.016)		(0.020)
Non-cognitive: TR			0.002	0.034
			(0.068)	(0.074)
Soc. Char	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes
Observations	1293	1021	1268	1015

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: SR denotes self-reported. TR denotes teacher-reported. Standard errors are clustered at the school level.

teacher-reported non-cognitive measure and column (4) adds both measures. The estimate for GSD remains stable and the measures for non-cognitive skills are not statistically significant for all specifications. The results hold when regressions are run for each student and teacher reported non-cognitive skill separately (not reported here). Once again, it does not seem that grade variance is associated with non-cognitive skills. The main inference from these estimates is that the association between grade variance and grade point average cannot be explained by non-cognitive skills.

In all three data sets, non-cognitive skills do not change the size or direction of the GSD estimate in the conditional correlation tables. There is no evidence that the association between grade variance and educational attainment can be explained by non-cognitive skills. As a result, even though the estimate between GSD and educational attainment is negative, there is no support of the non-cognitive skills hypothesis.

6 Conclusion

Throughout all explorations of the importance of the second moment of individual grade distribution, I find that individual grade variance is negatively associated with educational attainment. For both the United States and Norway, this association holds across the grade distribution and for both genders and estimates are robust to controlling for socioeconomic characteristics and school fixed effects. In addition, estimates remain negative when including measures of cognitive and non-cognitive skills. My results suggest that the negative association between grade variance and educational attainment is a general finding that is not country or context specific.

The cognitive-skill hypothesis is that high grade variance is associated with low educational attainment because it reflects low non-cognitive skills. This hypothesis is supported by the main results. However, the grade standard deviation estimate is larger in the up-

per end of the grading distribution for the United States and in the middle of the grading distribution for Norway, which does not support Lindqvist and Vestman (2011) who find that non-cognitive skills are more important in the lower end of the grading distribution. More importantly, using three different data sets, it is not possible to find a systematic relationship between non-cognitive skills and grade variance.

The other hypothesis is that high grade variance reflects being a specialist rather than a generalist, and that this is positively associated with educational attainment. However, the main results rather support the reversed generalist/specialist hypothesis, that it is beneficial to be a generalist. Why could it be beneficial to be a generalist? Lazear (2004) suggests that it might be beneficial to have a span of skills for certain studies or occupations. This might be the case also for higher education, which is often based on general knowledge, particularly in the United States where there is a long tradition for a liberal arts education in four-year colleges. Another possible explanation is that being a generalist increases your adaptability which could be beneficial for long-run outcomes (Hanushek, Woessmann, and Zhang, 2011). Testing these hypotheses is a topic for future research.

If institutions are interested in students with high ability and effort, but only use grade point average in the admission decision, they may not be accepting the best students. Students with low grade variance who are just below the grade point average cutoff are likely to outperform student just above the cutoff with high grade variance. My findings support that institutions should take grade variance, or other measures of skill, into account in admission decisions.

References

ACT (2015): "What is the ACT?," http://www.actstudent.org/faq/what.html, Reading date: 07.08.2015.

ASVAB (2015): "ASVAB Fact Sheet," http://official-asvab.com/docs/asvab_fact_sheet.pdf, Reading date: 07.08.2015.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. Ter Weel (2008): "The economics and psychology of personality traits," *Journal of Human Resources*, 43(4), 972–1059.

Carneiro, P., C. Crawford, and A. Goodman (2007): "The impact of early cognitive and non-cognitive skills on later outcomes," Discussion paper, CEE DP 92.

Clinedinst, M. E., and D. A. Hawkins (2011): "State of college admission," *Washington, DC: National Association for College Admission Counseling*.

Falch, T., O. H. Nyhus, and B. Strøm (2014): "Performance of Young Adults: The Importance of Different Skills," *CESifo Economic Studies*.

Hanushek, E. A., and D. D. Kimko (2000): "Schooling, labor-force quality, and the growth of nations," *American Economic Review*, pp. 1184–1208.

Hanushek, E. A., and L. Woessmann (2008): "The role of cognitive skills in economic development," *Journal of Economic Literature*, pp. 607–668.

Hanushek, E. A., L. Woessmann, and L. Zhang (2011): "General education, vocational education, and labor-market outcomes over the life-cycle," Discussion paper, National Bureau of Economic Research.

Haraldsvik, M. (2003): "Inntaksprosedyrer for den videregående skole: Grad av valgfrihet," Institutt for samfunnsøkonomi, NTNU.

Heckman, J. J. (1995): "Lessons from the bell curve," *Journal of Political Economy*, pp. 1091–1120.

Heckman, J. J., and T. Kautz (2013): "Fostering and measuring skills: Interventions that improve character and cognition," Discussion paper, National Bureau of Economic Research.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24(3), 411–482.

Hedges, L. V., and A. Nowell (1995): "Sex differences in mental test scores, variability, and numbers of high-scoring individuals," *Science*, 269(5220), 41–45.

Herrnstein, R. J., and C. Murray (2010): *Bell curve: Intelligence and class structure in American life*. Simon and Schuster.

Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014): "Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success," OECD Education Working Papers 110, OECD Publishing.

Lazear, E. P. (2004): "Balanced skills and entrepreneurship," *American Economic Review*, pp. 208–211.

Lindqvist, E., and R. Vestman (2011): "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment," *American Economic Journal: Applied Economics*, 3(1), 101–28.

Murnane, R. J., J. B. Willett, and F. Levy (1995): "The growing importance of cognitive skills in wage determination," *Review of Economics and Statistics*, 77(2), 251–266.

National Center for Education Statistics (2015): "Private School Enrollment," `http://nces.ed.gov/programs/coe/indicator_cgc.asp`, Reading date: 10.07.2015.

National Center for Research in Vocational Education and The Center for Human Resource Research, The Ohio State University (1984): *NLSY79 High School Transcript survey: Overview and Documentation*.

OECD (2015): "About PISA," `http://www.oecd.org/pisa/aboutpisa/`, Reading date: 07.08.2015.

Rosenberg, M. (1965): *Society and the adolescent self-image*. Princeton University Press Princeton, NJ.

Roth, B., N. Becker, S. Romeyke, S. Schäfer, F. Domnick, and F. M. Spinath (2015): "Intelligence and school grades: A meta-analysis," *Intelligence*, 53, 118 – 137.

Rotter, J. B. (1966): "Generalized expectancies for internal versus external control of reinforcement.," *Psychological monographs: General and applied*, 80(1), 1.

SAT (2015): "About the SAT," `https://sat.collegeboard.org/about-tests/sat`, Reading date: 07.08.2015.

Segal, C. (2012): "Working when no one is watching: Motivation, test scores, and economic success," *Management Science*, 58(8), 1438–1457.

Stevenson, H. W., and R. Nerison-Low (2002): "To Sum It Up: Case Studies of Education in Germany, Japan, and the United States.," .

The Norwegian Directorate for Education and Training (2015): "Grunnskolens informasjonssystem," https://gsi.udir.no/application/main.jsp?languageId=1, Reading date: 29.12.2015.

U.S. Bureau of Labor Statistics (2015): "NLSY79 Appendix 21: Attitudinal Scales," https://www.nlsinfo.org/content/cohorts/nlsy79/other-documentation/codebook-supplement/nlsy79-appendix-21-attitudinal-scales, Reading date: 10.07.2015.

A National Longitudinal Survey of Youth, 1979

A.1 Data description

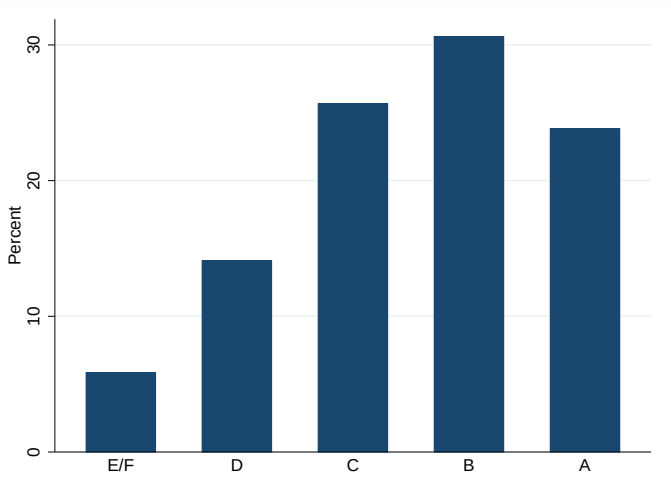
Between 1980 and 1983, transcript information was collected for respondents who were 17 years of age or older and expected to complete high school in the United States. The data include up to 64 courses. Of the 6,111 respondents, 5,009 have non-missing transcript data (see Table A4). Information for each course on the transcript includes (1) grade level for which the course was taken (2) a code for the high school course (3) the final or computed grade for that course (4) the source for the final grade and (5) the credits received. Courses are divided into 22 subject areas, listed in Table A1. For a complete list of course codes, see (National Center for Research in Vocational Education and The Center for Human Resource Research, The Ohio State University, 1984). Students either receive a pass/fail grade, or grades A-F, where grade A is coded to value 4, grade B to value 3, grade C to value 2, grade D to value 1. Grade F is a failing grade and is coded to value 0. Figure A1 shows the distribution of grades for the 214,507 grades in the sample. The analysis is restricted to students with 10 or more valid grades. Figure A2 shows the distribution of number of grades in the sample. The data reduction is presented in Table B1.

Table A1: NLSY79 - Course subject area in transcript data

	N	Percent
Agriculture	1718	0.79
Art	7405	3.40
Business	3058	1.40
Distributive education	1038	0.48
English	43119	19.80
Foreign Language	7830	3.59
Health occupations education	294	0.13
Health and physical education	25129	11.54
Home economics	9707	4.46
Industrial arts	7390	3.39
Mathematics	23496	10.79
Music	6517	2.99
Natural sciences	19926	9.15
Office occupations education	11287	5.18
Social studies	34354	15.77
Technical education	62	0.03
Vocational	2971	1.36
Safety and driver education	3827	1.76
Junior ROTC	450	0.21
Philosophy and religion	1500	0.69
Study skills	731	0.34
Career education	4120	1.89
Missing	1875	0.86
Total	217804	100.00

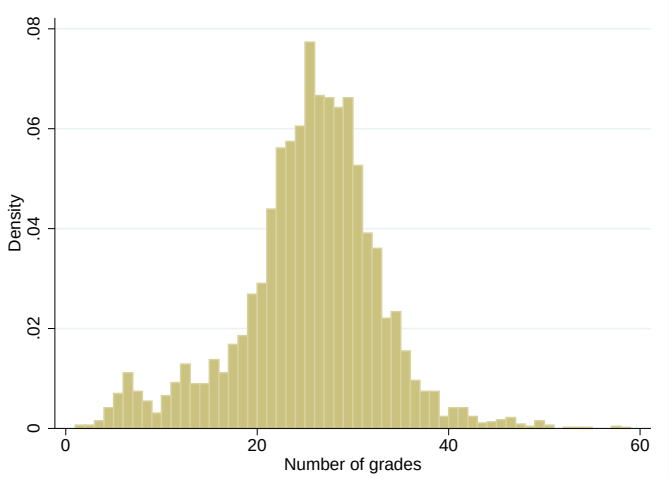
Note: Missing denotes missing course code but non-missing course grade. See National Center for Research in Vocational Education and The Center for Human Resource Research, The Ohio State University (1984) for a detailed list of the course codes.

Figure A1: NLSY79 Grade distribution



Note: The figure includes 113,213 grades ranging from E/F (non-pass, lowest) to A (highest) for 4389 students from the NLSY79 survey. The sample includes students from the main sample with 10 or more valid grades and with non-missing educational attainment at age 30.

Figure A2: NLSY79 - Number of grades, grades 9-12



Note: The figure includes 4,577 students from the NLSY79 survey who have grades reported transcript data, are from the main sample and have non-missing educational attainment at age 30. 205 students have less than 10 grades, and are dropped in the analysis. The final sample is thus 4,389 students (see Table A4).

A.2 Cognitive and non-cognitive skills

Table A2: The NLSY79 Rotter – Locus of control questions

1a	What happens to me is my own doing.
1b	Sometimes I feel that I don't have enough control over the direction my life is taking.
2a	When I make plans, I am almost certain that I can make them work.
2b	When I make plans, it is not always wise to plan too far ahead, because many things turn out to be a matter of good or bad fortune anyhow.
3a	Getting what I want has little or nothing to do with luck.
3b	Many times we might just as well decide what to do by flipping a coin
4a	Many times I feel that I have little influence over the things that happen to me.
4b	It is impossible for me to believe that chance or luck plays an important role in my life.

Note: The Rotter Locus of Control Scale is a four item forced choice questionnaire and is an abbreviated version of the 60-item Rotter scale. Scores are generated for each pair of items. Internal control: Much closer=1 Slightly closer =2 External control: Much closer=3 Slightly closer=4. Scores of 4 pairs were summed. Total score could range from 4 to 16 points. If one item is missing, the scale score is coded as missing(U.S. Bureau of Labor Statistics, 2015). In this paper, scores are reversed such that a higher score is more internal control, and thus reflects higher non-cognitive skills (values from 0 to 12). The test was administered in the NLSY79 in 1979.

Table A3: The NLSY79 Rosenberg Self-Esteem Scale questions

1	I am a person of worth.
2	I have a number of good qualities.
3	I am inclined to feel that I am a failure.
4	I am able to do things as well as most other people.
5	I felt I do not have much to be proud of.
6	I take a positive attitude toward myself.
7	I am satisfied with myself.
8	I wish I could have more respect for myself.
9	I certainly feel useless at times.
10	At times I think I am no good at all.

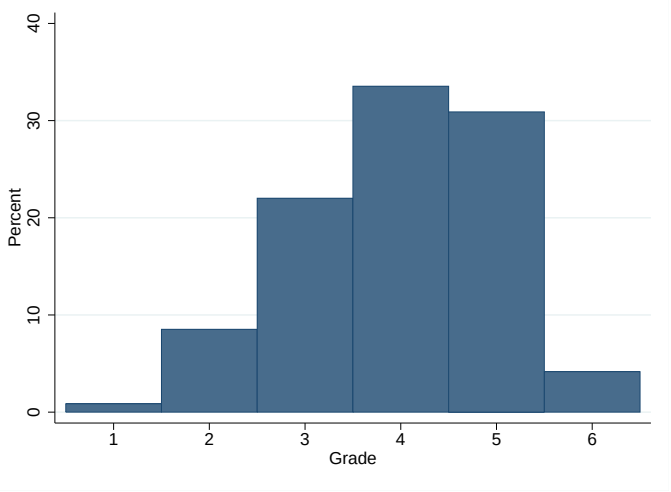
Note: The scale contains 10 statements about self-approval and disapproval to which the respondents are asked to strongly agree, agree, disagree or strongly disagree. Higher scores are associated with higher self-esteem. Scoring for items 3, 5, 8, 9, 10: strongly agree=0 agree=1 disagree=2 strongly disagree=3. Scoring for items 1, 2, 4, 6, 7 is reversed so that a higher score indicates higher self-esteem. Scores of 10 items were summed. Total score could range from 0 to 30 points. If one item is missing, the scale score is coded as missing(U.S. Bureau of Labor Statistics, 2015). The test was administered in the NLSY79 in 1979.

Table A4: Data Reduction NLSY79

	N	Reduction	% Reduction
1. Complete sample	12686		
2. Main sample	6111	6575	51.83 %
3. Non-missing transcript data	5009	1102	18.03 %
4. Non-missing educational outcome	4577	432	8.62 %
5. 10 or more valid grades	4389	188	4.11 %
6. Non-missing cognitive skills	4243	146	3.33 %
6. Non-missing non-cognitive skills	4226	163	3.71 %
6. Non-missing cognitive and non-cognitive skills	4136	253	5.76 %

B Norwegian register data

Figure B1: NRD



Note: 2,037,789 grades ranging from 1 (lowest) to 6 (highest) for 158,308 students leaving lower secondary education 2002-2004. About 90% of students have 13 valid grades.

Table B1: Data Reduction NRD

	N	Reduction	% Reduction
1. Sample 2002-2004	168,151		
2. 10 or more valid grades	162,831	5,320	3.16 %
3. 16 years old	159,077	3,754	2.31 %
4. Non-missing school information	158,308	769	0.48 %

Note: Restriction number 3 is that the student has to be 16 years old when graduating from lower secondary education.

Table B2: NRD: Academic track - cognitive and non-cognitive skills

	(1) ACA	(2) ACA	(3) ACA	(4) ACA
GPA	0.244***	0.233***	0.291***	0.286***
	(0.002)	(0.003)	(0.003)	(0.005)
GSD	-0.018***	-0.017***	-0.009***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)
Cognitive		0.012***		0.005
		(0.003)		(0.003)
Non-cognitive			-0.049***	-0.048***
			(0.003)	(0.003)
Soc. Char	Yes	Yes	Yes	Yes
CohortxSchool FE	Yes	Yes	Yes	Yes
R-squared	0.301	0.301	0.303	0.302
N	158,308	158,289	158,308	158,289
Number of groups	3,397	3,397	3,397	3,397

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the school level.

Table B3: Vocational graduate - cognitive and non-cognitive skills

	(1) VOC graduate	(2) VOC graduate	(3) VOC graduate	(4) VOC graduate
GPA	0.270***	0.205***	0.179***	0.096***
	(0.002)	(0.004)	(0.004)	(0.006)
GSD	-0.000	0.009***	-0.023***	-0.014***
	(0.002)	(0.002)	(0.002)	(0.002)
Cognitive		0.076***		0.087***
		(0.004)		(0.004)
Non-cognitive			0.091***	0.099***
			(0.004)	(0.004)
Soc. Char	Yes	Yes	Yes	Yes
CohortxSchool FE	Yes	Yes	Yes	Yes
R-squared	0.242	0.246	0.248	0.253
N	80,725	80,710	80,725	80,710
Number of groups	3,306	3,306	3,306	3,306

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the school level.

Table B4: Academic graduate - cognitive and non-cognitive skills

	(1) ACA graduate	(2) ACA graduate	(3) ACA graduate	(4) ACA graduate
GPA	0.210***	0.174***	0.178***	0.130***
	(0.003)	(0.004)	(0.004)	(0.005)
GSD	-0.013***	-0.009***	-0.016***	-0.012***
	(0.002)	(0.002)	(0.002)	(0.002)
Cognitive		0.038***		0.046***
		(0.004)		(0.004)
Non-cognitive			0.036***	0.043***
			(0.003)	(0.003)
Soc. Char	Yes	Yes	Yes	Yes
CohortxSchool FE	Yes	Yes	Yes	Yes
R-squared	0.223	0.225	0.225	0.228
N	72,839	72,838	72,839	72,838
Number of groups	3,194	3,194	3,194	3,194

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the school level.

Table B5: Upper secondary education GPA - cognitive and non-cognitive skills

	(1) GPA USE	(2) GPA USE	(3) GPA USE	(4) GPA USE
GPA	1.012***	0.756***	1.152***	0.881***
	(0.006)	(0.008)	(0.008)	(0.011)
GSD	-0.012***	0.016***	0.004	0.026***
	(0.003)	(0.003)	(0.003)	(0.003)
Cognitive		0.280***		0.261***
		(0.007)		(0.007)
Non-cognitive			-0.162***	-0.125***
			(0.007)	(0.007)
Soc. Char	Yes	Yes	Yes	Yes
CohortxSchool FE	Yes	Yes	Yes	Yes
R-squared	0.542	0.554	0.546	0.557
N	83,740	83,737	83,740	83,737
Number of groups	3,208	3,208	3,208	3,208

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the school level.

Table B6: Started higher education- cognitive and non-cognitive skills

	(1) Started HE	(2) Started HE	(3) Started HE	(4) Started HE
GPA	0.137***	0.118***	0.162***	0.145***
	(0.003)	(0.004)	(0.004)	(0.005)
GSD	-0.004**	-0.002	-0.002	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Cognitive		0.020***		0.016***
		(0.003)		(0.003)
Non-cognitive			-0.029***	-0.027***
			(0.003)	(0.003)
Soc. Char	Yes	Yes	Yes	Yes
CohortxSchool FE	Yes	Yes	Yes	Yes
R-squared	0.111	0.112	0.113	0.113
N	83,740	83,737	83,740	83,737
Number of groups	3,208	3,208	3,208	3,208

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Standard errors are clustered at the school level.

C Data from the Development in Adolescence Project

The Development in Adolescence Project (CDAP) is a longitudinal survey of 1559 middle school students and their teachers from 8 different schools. The same students receive a survey in four rounds, the fall and spring of eighth grade and the fall and spring of ninth grade. Their teachers in math, science, English and social studies also receive a survey in each round. The data also include grades from math, science, English and social studies for each semester. I use data from rounds 1 and 2. Only students with one or no missing grades are included in the analysis. Two schools are dropped from the analysis, one due to missing grade data and another due to different grading practices. This leaves a sample

of 1293 students.

Grade point average (GPA) is calculated as the average of all grades received during the two rounds. Grade standard deviation (GSD), used as a measure of grade variance, is calculated as the standard deviation of the same grades used to calculate grade point average. GPA and GSD are then standardized for the whole sample. Socioeconomic characteristics include gender, ethnicity (dummy variables for Hispanic, Asian, African American, multiethnic or other) birth date, being an English language learner, receiving reduced/free lunch and receiving special education. Rather than exclude students with missing values on control variables, dummy variables for missing are constructed and included in the regressions. Descriptive statistics for GPA, GSD and socioeconomic characteristics are listed in Table C1.

Students' self-reported non-cognitive skills in each round include, among other things, (1) delay discounting, (2) grit, (3) self-control: work, (4) self-control: interpersonal, (5) gratitude, (6) actively open-minded thinking, (7) prosocial purpose and (8) internal locus of control. To create a joint measure of students' non-cognitive skills, each measure is standardized with mean 0 and standard deviation 1 before standardizing the sum of these measures with mean 0 and standard deviation 1. There are 272 students with missing information on one or more measures, reducing the sample to 1021. Teacher-reported non-cognitive skills for individual students in each round include (1) grit, (2) self-control: work, (3) self-control: interpersonal, (4) gratitude, (5) actively open-minded thinking and (6) prosocial purpose.

Teacher self-reported measures are averages across all teachers for each student. To create a joint measure of teacher-reported non-cognitive skills, each measure is standardized with mean 0 and standard deviation 1 before standardizing the sum of these measures with mean 0 and standard deviation 1. There are 25 students with missing information on one or more teacher-reported measures, reducing the sample to 1268. Descriptive statistics for student self-reported and teacher-reported non-cognitive skills are listed in Table C1.

Table C1: Development in Adolescence Project - Descriptive statistics

	Total		Boy		Girl	
	mean	(sd)	mean	(sd)	mean	(sd)
Girl	0.49	(0.50)	0	(0)	1	(0)
Hispanic	0.16	(0.37)	0.16	(0.37)	0.17	(0.37)
Asian	0.11	(0.32)	0.12	(0.33)	0.11	(0.31)
Multiethnic or other	0.0085	(0.092)	0.012	(0.11)	0.0047	(0.069)
African American	0.48	(0.50)	0.47	(0.50)	0.49	(0.50)
Birth month	6.68	(3.48)	6.70	(3.48)	6.66	(3.48)
English language learner	0.14	(0.35)	0.14	(0.35)	0.14	(0.35)
Special education	0.16	(0.36)	0.20	(0.40)	0.11	(0.32)
Free/reduced lunch	0.66	(0.47)	0.64	(0.48)	0.68	(0.47)
Non-cognitive: self-reported	0	(1.00)	-0.022	(1.01)	0.022	(0.99)
Non-cognitive: teacher reported	0	(1.00)	-0.22	(1.02)	0.23	(0.93)

Note: N=1293, with 659 boys and 634 girls. For Non-cognitive: self-reported, N=1021, with 514 boys and 507 girls. For Non-cognitive: teacher reported, N=1268, with 650 boys and 618 girls.