Four Essays on the Determinants of Human Capital Accumulation in Norway

Leroy Andersland

Thesis for the Degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2017

UNIVERSITY OF BERGEN

2017

Four Essays on the Determinants of Human Capital Accumulation in Norway • Leroy Andersland

NORDIC ECOLABEL 241 PRINTED MATTER 731

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 978-82-308-3856-3

Four Essays on the Determinants of Human Capital Accumulation in Norway

Leroy Andersland

Thesis for the Degree of Philosophiae Doctor (PhD)
at the University of Bergen

2017

Date of defence: 19.01.2018

Year: 2017

Title: Four Essays on the Determinants of Human Capital Accumulation in Norway

Name: Leroy Andersland

Print: Skipnes Kommunikasjon / University of Bergen

For my family and friends.

Acknowledgements

I thank my supervisors Katrine V. Løken and Espen Bratberg for guiding my work on this dissertation; the results would not have been the same without their thoughtful guidance. My main supervisor Katrine has been very helpful in developing the ideas and the methodological framework of the analysis. Espen contributed to the methodological framework and by helping me understand the direction in which I should lead the work.

I also thank my co-author Øivind A. Nilsen who has been both a mentor and a co-author on the article on which we collaborated. Sergio Firpo, Magne Mogstad and Kjell G. Salvanes have provided important guidance.

My fellow PhD students and the postdoctoral fellows at the places I have studied have been both academic and non-academic sources of inspiration. A special thanks goes to Sigurd Birkeland, Manudeep Bhuller, Erlend Bøe, Inger Sommerfeldt Ervik, Cathrin Fløgstad, Ingvil Gaarder, Julian Vedeler Johnsen, Andreas Kostøl, Sturla Løkken, Andreas Steinmayr, and Hans Martin Straume.

My research stays would not have been as productive without the help of Rolf Aaberge, Prashant Bharadwaj, Gordon Dahl, and Ariel Kalil. The Meltzer Foundation, Trygdegruppen, and SV-fakultet have provided additional funding.

Oslo, August 2017

Leroy Andersland

Abstract

This thesis consists of four articles that use administrative data to explore the Norwegian education system, from childcare to high school. The goal of the dissertation as a whole is to uncover and quantify the impact of mechanisms that explain why some students prosper, while others do not. The first article seeks to determine whether teachers discriminate against students based on observable characteristics of the students that are available to us. Previous literature has provided inconclusive results on whether, for example, males and immigrants are discriminated against in Norway and Sweden. The standard procedure has been to compare grades awarded by the student's teacher (non-blind) with grades awarded by examiners who do not know the identity of the student (blind). This article makes three main contributions to the literature. First, it provides a coherent econometric framework in which to study grade discrimination in schools. Secondly, results are presented from several different types of data sets, which helps determine the underlying structure that determine teacher grading. Lastly, evidence is presented implying that an adjustment should be made when comparing non-blind and blind grading, since the scores are not directly comparable. This correction would partly reconcile some of the previous contradictory findings in the literature. Generally, this correction takes into account the fact that discrimination against students belonging to lower scoring groups is often more significant when holding ability fixed.

The second article explores the nature of peer effects in high school. Peer effects can be defined as a composite of factors that explain why interacting with peers with certain behaviors or characteristics affects a student's own behavior. In 2005, the city of Bergen was exposed to a reform that changed the high school intake system from a geographical, neighborhood-based intake system to a GPA-based intake system. The reform greatly altered the composition of peer characteristics for comparable students before and after the reform.

The reform led to a greater degree of student tracking; low-achieving students received lower variation in peer characteristics and low achieving peers, while high-achieving students reviewed lower variation in peer characteristics and higher achieving peers. In line with recent findings from field experiments, we find that students from all ability levels gained from this reform, with low-ability students appearing to gain the most. This article makes three main contributions to the literature. First, we use a new type of natural experiment to explore peer effects. Earlier natural experiments included voucher lotteries, desegregation programs, and high school acceptance limits. Secondly, this article contains an analysis of tracking using a natural experiment. Our results are directly relevant for policymakers trying to determine what type of intake system to use. Third, we present evidence on several interesting high school outcomes such as grades, exam scores, and absence rates.

The third article explores the effect of attending childcare on children. In recent years, formal childcare has become the dominant mode of care for children aged 1–5 in Norway. Yet, the effect of attending public childcare on different groups of children is still not well understood. This article employs the significant capacity buildup of the Norwegian childcare sector in the 2000s to explore the effect of formal childcare. The findings suggest that the effects of formal childcare are heterogeneous. We find no average effect of the expansion, while we do find positive effects in municipalities with high childcare quality, and negative effects in municipalities with low childcare quality. The analysis reveals that the reform mostly affects children aged 3–5 in municipalities with high childcare quality, while it affects children aged 1–2 in municipalities with low childcare quality. In addition, positive effects seem to be driven by children of high socioeconomic status, while negative effects are stronger for children of low socioeconomic status. This article contributes with an analysis of a recent expansion in universal childcare in Norway. The results confirm previous findings of positive effects for 3–5 year old children in formal childcare using a new natural experiment and data set, and different outcomes. Furthermore, we employ a novel identification strategy in our analysis, leveraging the fact that the expansion was more comprehensive in municipalities with lower childcare coverage to begin with. Moreover, the results reveal heterogeneity that seem to be important to understand how different groups of children are affected by public childcare. Younger children seem less likely to gain from childcare. The findings show negative effects for children of low socioeconomic status in low-quality municipalities, and no effects for these children in high-quality municipalities. This adds to the discussion on how public institutions affect intergenerational transmission of inequality in outcomes.

The fourth article explores how the care of children is affected by a reform that increased the price of formal childcare. To gain a deeper understanding of why some children benefit from formal childcare while others do not, it is important to have detailed information on what the alternative mode of care is for different types of children. The Cash-for-Care benefit was introduced in 1998 and provides funds to parents who do not send their 1–2 year old children to formal childcare. We find that for the households that are affected by this reform, the main alternative mode of care is parental care. The main alternative for households of low socioeconomic status is parental/relative care, whereas the main alternatives for families of high socioeconomic status include day parks and nannies as well as parental care. The analysis also reveals that care decisions for young children change due to price changes in formal childcare, with point estimates of price elasticities of -0.33 and -0.25. This article focuses more closely on the effects of the CFC reform on children than does previous literature. The analysis uses survey data that allow for a detailed inspection of responses to the reform. Moreover, the survey data are compared to administrative data to verify the results.

Contents

1 Introduction

This dissertation contains four empirical articles exploring the determinants of human capital accumulation in Norway. The first article compares teacher-given (non-blind) grades and externally and anonymously graded (blind) grades to examine the extent of discrimination in schools in Norway. Its primary contribution is to provide evidence on the relationship between non-blind and blind grades with subject ability, and evaluate the consequences of deviations in this relationship between the two grades. The second article examines the high school intake system in a particular municipality in Norway, Bergen, to explore how a change in peer characteristics influences student outcomes. We add to the literature with an examination of a natural experiment that allows us to explore the effects of dividing students into groups based on prior ability across high schools. The third article uses a capacity expansion of childcare in Norway to study the effect of attending childcare facilities on later test scores. We employ a novel identification strategy for this question, using pre-reform coverage rates and studying in particular childcare quality and heterogeneity by child age. The fourth article explores household responses to a reform that changed the price of formal childcare. In contrast to the current literature, we focus on the effects of this price change on the care arrangements for children rather than on parents' labor market outcomes.

The articles share certain common elements. Firstly, they all take advantage of high-quality administrative registry data: Norwegian registry data received the highest ratings in a study carried out by Atkinson, Rainwater, & Smeeding (1995). Secondly, all the articles seek to identify causal effects and exploit natural experiments for identification. Natural experiments are different from laboratory or field experiments in that they involve contexts that are generated by a reform, policy change, rule, or natural disaster rather than by the researcher(s).

While each analysis is set in a specific circumstance, time, and place, they all aim to contribute to the general knowledge of the nature of human capital accumulation. Quantifying the contribution of different factors is a task that is developing rapidly as empirical methods, economic theory, data quality, and concepts evolve.

The next sections contain summaries of each article in the dissertation, followed by a discussion of the empirical strategy used in the analysis.

1.1 The Extent of Biased Grading at School

The first article is part of a growing literature that explores discrimination in the education system. The analysis of discrimination in schools is seen as an important area alongside discrimination in the labor market (hiring, wages), housing, and law enforcement (policing, judges, lawyers). Becker (1957) developed a theory of discrimination based on the concept of "taste for discrimination." This occurs when an agent discriminates a group because he has a disutility associated with that group. In the labor market context, the standard example often referred to is the case when employers dislike working with people from a particular group, and is willing to pay a higher wage to employ a person outside that group with equal productive attributes. The model provides predictions for firm performance and wage differentials under various conditions.

The concept of taste for discrimination can readily be applied to the school-grading context. Examples where teachers base grading on group membership or student characteristics other than objective attributes that are supposed to be included in the student's grade can occur because of teachers' preferences. Teachers then grade because they like a group of students better than another group, or that they do not agree on the common set of course objectives, and grade based on other attributes of the student.

An alternative theory explaining why discrimination occurs is often referred to as "statistical discrimination" (Phelps, 1972; Arrow, 1973). In this model, the discriminating agent gets a noisy signal about another agent's ability, while at the same time having prior information on the average ability of different groups (Aigner & Cain, 1977). The discriminating agent then bases his or her decision on both the noisy signal and the prior information on group averages to make a decision. If females are more productive on average, and an employer is supposed to hire a job applicant based on one interview, the employer is using information both from the interview of both the male and female, and prior information about the higher female group averages to decide whom to hire. If the male and female perform equally well in the interview, the female is hired because of the higher average productivity of females. A similarity can be drawn to the school-grading context, when the teacher is supposed to give a grade based on a course or exam performance. This framework implies that teachers use both observed performance of the student and other characteristics observable to the teachers (group means, for example) to set the grade.

The theories of taste-based discrimination and statistical discrimination can explain why teachers discriminate in grading. If teachers engage in taste-based discrimination, measures to reduce teachers' ability to perform discretionary grading would lead to a reduction in this type of bias. For example, being more specific about what should be included in the grade. Implementing this measure would not necessarily improve the situation if the discrimination were explained by statistical discrimination. In this case, implementing measures to reduce the noise in grading would help reduce discrimination of all individuals.

In some contexts, to measure the extent of discrimination is to identify specific characteristics and determine if that is used to discriminate. For example, in hiring decisions, holding all other characteristics fixed, how much more/less likely is an immigrant to be hired? The reason for this is that the law states that employers are not allowed to discriminate based on certain characteristics. Employers are of course still allowed to discriminate based on other traits that also are not necessarily directly job-related. It is possible to argue that, in the school setting, this analytical approach to measuring discrimination is less appealing. There are two reasons for this. First, being a member of a specific group holding all other characteristics fixed is an abstract exercise and may have little relevance to how discrimination works outside randomized controlled experiments. This is an argument that can also be made against measuring this type of discrimination in labor market contexts. Second, discrimination in school will manifest in the grades that students receive. All types of discrimination may therefore explain outcome differences between groups. For example, if employers and teachers do not discriminate based on gender, but boys behave worse than girls, then taking the bad behavior into account would not be considered discrimination in hiring settings since employers can take this into account even if it does not affect the ability to do tasks. However, if teachers based grading on behavior when behavior is not supposed to be included, we argue that this bias is worth measuring.

1.1.1 Measuring School Discrimination Using Administrative Data

The goal of this paper is to assess whether differences between assessments by the student's own teacher (non-blind scores), and tests graded anonymously (blind scores), can be interpreted as discrimination by teachers. We focus on two types of data generating processes of the blind and non-blind scores. The first type occurs when the student's own teacher and another examiner are marking the same exam. As in most previous studies, the second is a data-generating process in which the student's own teacher and an external teacher are marking different tests that are meant to measure the student's knowledge of the same material. We present a parsimonious econometric framework that shows, for each data-generating process, the assumptions under which one can identify bias in teachers' assessment from a comparison of blind and non-blind test scores. This framework lays the groundwork for our empirical analysis, where we use data from the Norwegian school system to estimate and interpret differences between non-blind and blind assessment of students.

The literature that compares non-blind and blind evaluations of students' performance begins with Lavy (2008). The study tests for gender stereotyping in Israeli high schools by comparing grades given by teachers that know the students (non-blind) to grades of teachers that do not know the students (blind) of two exams that test the same skills. Using a difference-in-difference (DD) design, the study found evidence of a bias against male students. This finding has been confirmed in studies from other countries. Lindahl (2007) compared the non-blind assessments and blind test score evaluations of Swedish students, and found the same gender difference as well as a difference favoring non-native students. Falch and Naper (2013) found the same pattern at the end of lower secondary school in Norway.

In these studies, non-blind and blind evaluations were not of the same test. The findings suggesting positive discrimination of females may actually suggest that there is something else that is the reason for this difference. Hinnerich et al. (2011, 2015) collected data that allow comparisons of non-blind and blind evaluations of the same exam in Swedish schools. In these studies, an external teacher that does not know the student grades the same exam as a teacher that knows the student. The main findings from the two studies is that, even though local teachers raised grades on average, the results do not suggest the existence of any gender bias, while they find evidence of discrimination against students with foreign backgrounds.

We call datasets that include a non-blind and blind score of two tests that are meant to test the same skill of the student administrative datasets. Datasets with non-blind and blind evaluations of the same test we call non-administrative. Writing out a model for grades in administrative data lets us clearly discuss the content of grade differences. Let us assume that $\tilde{Y}_{is}^{n}$, the grade given by the teacher in the administrative data, can be written as

$$\tilde{Y}_{is}^{n} = t(X_{is}) + \tilde{t}(X_{is}) + \rho\theta_{is} + (1-\rho)\tilde{\theta}_{is} + \tilde{\varepsilon}_{is}^{n}$$

$$t(X_{is}) = t(G_{is}, \kappa_{is}) = \alpha + \beta G_{is} + \gamma\kappa_{is}$$

$t(X_{is})$ is the biased grading function, or simply the bias. The function $\tilde{t}(X_{is})$ explains why some students perform relatively better under in-class tests graded by the teacher. The variable X_{is} is a vector that contains G_{is}, which are some observable characteristics to the researcher and the teacher. κ_{is} represents student behavior in class, and $\tilde{\theta}_{is}$ is a compound of other information about the students that the teacher use to grade. In particular, $\tilde{\theta}_{is}$ is, for example, other student abilities/behavior, grades in other subjects, or previous grades. Importantly, in administrative data, $\tilde{\theta}_{is}$ could also include other subject skills not tested in the blind test. κ_{is} and $\tilde{\theta}_{is}$ are not necessarily observable to the researcher. The variable θ_{is} is the ability being measured in the external test. The parameter ρ is the fraction of that ability that is measured by the teacher, or reflects the difference in learning goals weighting between internal and external examiners. $\tilde{\varepsilon}_{is}^{n}$ is an idiosyncratic error. The parameter α captures grade inflation, and β captures discrimination in favor of a group of students with observable characteristics G_{is} , while γ and $(1-\rho)$ capture the effects of components that are unobservable to us but that are used by the teacher when grading exams.

The grade given on the exam by the external grader is Y_{is}^{b}

$$Y_{is}^{b} = \theta_{is} + \varepsilon_{is}^{b}.$$

The grade difference can then be written as:

$$\Delta_{is}^{o} = \tilde{Y}_{is}^{n} - Y_{is}^{b} = t(X_{is}) + \tilde{t}(X_{is}) + (\rho - 1)\theta_{is} + (1-\rho)\tilde{\theta}_{is} + \tilde{\varepsilon}_{is}^{n} - \varepsilon_{is}^{b}.$$

The parameter ρ represents the difference in the relationship between subject ability and non-blind and subject ability and blind. The literature that discusses structural parameters in the estimation of bias is concerned with the size of ρ. This is because it is seen as an indicator for whether two tests used to measure discrimination measure the same skills. For example, Terrier (2016) uses an instrumental variable strategy for French data and cannot reject that ρ is 1. Estimating the size of ρ is important in determining the size of bias conditioning on subject ability. According to our model, the grade difference will then be a function of subject ability if rho is different from 1. To estimate the size of ρ, we use a grouping strategy (Deaton, 1985) by regressing the grade difference on grouped blind score averages. Our results using this strategy suggest that ρ tends to be below 1. We find evidence of this both when two teachers grade different exams and when teachers that know the student, and teachers that do not, grade the same exam. Finding a ρ below 1 in the administrative data may indicate that the two tests actually measure different skills. However, finding such a relationship in the non-administrative data suggests that there also is another explanation. For example, the students' teacher has additional information or face different incentives.

Our model emphasizes the importance of $\tilde{\theta}_{is}$. If the other information used by teachers in grading is subject skills not tested in exams, grade differences could reflect this. When one uses administrative data where the non-blind to blind grade differences come from different tests, it is important that one use tests that are meant to test the same skills. For example, teacher assessments normally cover more material, and one group could be better at one part of the material one year. By using recordings from several years, this should not matter if the material in exams changes year to year to cover all learning goals in a subject. In our project, we have been careful to use teacher assessments in subjects where oral performance does not count. For instance, we did not use recordings of grade differences in English since, in this subject, there is an oral component in the learning goals, and there is not a separate oral teacher assessment grade in English. The model also emphasizes the possible content of $\tilde{t}(X_{is})$. If groups of students perform differently under different types of exams, grade differences between groups could be due to this phenomenon, and not bias.

To evaluate the importance of these issues in our administrative data, we use data from trials where local and external teachers grade the same exam. In this case, the fact that one group of students performs better under one test type cannot explain group differences. In addition, the problem of two tests testing different subject skills disappears. For the administrative sample we study, our results do not suggest that these factors explain the group differences. However, further analysis should work to obtain additional data to increase the precision of the estimates.

Lastly, our model makes it possible to discuss how to interpret discrimination estimates. Discrimination, or stereotyping, can be seen as the bias one group receives compared to another, holding all other factors constant. How do teachers' gender stereotypes affect grading? Our model makes it clear that this is not possible to measure using the audit data we have available. This is because other student characteristics that teachers use in grading, κ_{is} and $\tilde{\theta}_{is}$, can be correlated with groups. For example, male students could behave worse than female students. To measure discrimination based on student characteristics, holding all other factors fixed, one needs to randomize student characteristics, as done by Bertrand and Mullainathan (2004) and Hanna and Linden (2009).

The consequence of taking into account that rho is less than 1 is that it reveals that the gender bias holding ability level constant is somewhat larger than not holding the ability level constant. Furthermore, when estimating bias between groups that have larger ability differences, the estimate of group bias changes even more when holding the ability level constant.

1.2 Peer Effects from a School Choice Reform

The question of how peer effects operate remains unsettled. When referring to peer effects in school, one is typically interested in how the ability of an individual's peers influence that individual's outcomes. There are multiple motivations behind this interest. For example, school administrators or policymakers may have an incentive to organize students within or across schools to achieve the best learning outcomes for all, such as increasing the mean outcome. The existence of peer effects then becomes an important part of the decision on how to group students within or across schools. Furthermore, parents of school-age children have an interest in knowing what environment most enriches their child's learning experience. It is difficult for parents to choose specific study partners for their children, but they can influence it by deciding the type of school to which they will send their children. The existence of peer effects will have greatest consequences for the individual student. For example, the influence of a high- versus low- ability study partner can significantly affect an individual's future outcome. Lastly, researchers want to learn about peer effects as one of many components that can explain why some students prosper, while others do not.

This article contributes to the literature that uses natural experiments to examine peer effects. The types of natural experiments that have been used earlier include housing vouchers (Kling, Ludwig, & Katz, 2005; Kling, Liebman, & Katz, 2007; Ludwig et al., 2013; Chetty, Hendren, & Katz, 2015), busing students (Angrist & Lang, 2004), and school assignment lotteries (Clark, 2010; Jackson, 2013; Abdulkadiroğlu, Angrist, & Pathak 2014). This article contributes by using a school choice reform to study peer effects. In 2005, the norwegian city of Bergen experienced a reform that changed the composition of students at different high schools. The reform changed the intake system from a catchment area approach to a performance-based intake system. One consequence of implementing the new system was the concentration of high-ability students at certain high schools in the central area of Bergen.

Since we aim to keep school type and travel distance fixed, the identification of peer effects is based on comparing the outcomes of the same type of students that attended these downtown schools before and after school choice reform to comparable students in other cities. The group of students that attend these attractive schools are high-ability students who reside in the downtown area, where high ability is defined as those scoring in the top 25% of their cohort in a city. They attended the downtown schools before reform because they lived in those schools' catchment areas and they attend downtown schools afterward because they still have that option, and there are few reasons for them to increase travel time to attend a school in the suburbs. The study shows that high-ability students in downtown Bergen attend schools with students that had on average 0.65 standard deviation (SD) higher middle school GPAs after reform. This is equivalent to moving from the median school to a school among the top 10% of pre-reform schools in Bergen and comparison cities. The evidence suggests that some exam scores increase as a consequence of the reform for this group of students.

For lower-ability students, the reform implied attending high school with less variation in peer achievement. Consistent with recent findings (Boiji et al. 2017, Carrell et al. 2013, Duflo et al. 2011), our results suggest that high school performance for these students increased as a consequence of the reform. The intake reform led to a natural experiment that generated a type of tracking similar to that achieved in experiments. The reform makes it possible to identify effects on high ability students of changing peers from mixed to high ability (high-high). For low ability students it is possible to find effects of changing peers from mixed to low (low-low). The effects of this type of tracking are relevant in cases where one decides between dividing a group based on prior ability or not.

1.2.1 School and peer effects

There is an extensive literature from the US on the effect of attending Catholic high schools (Coleman, Hoffer, & Kilgore, 1982; Bryk, Lee, & Holland, 1993; Evans & Schwab, 1995; Figlio & Stone, 1999; Grogger et al., 2000; Altonji, Elder, & Taber, 2005), and more recently the effect of attending charter schools (Hoxby & Rockoff, 2005; Hoxby, Murarka, & Kang, 2009; Gleason et al., 2010; Abdulkadiroğlu et al., 2011; Dobbie & Fryer, 2011, 2015; Angrist et al., 2016). The results from studies of Catholic schools show positive effects, while there are emerging results of clear positive short and long run effects of attending some types of charter schools.

Even though disentangling school effects from peer effects is not the main focus of these studies, it remains an important issue. It is likely that good schools attract ambitious and high-achieving students. At the same time, it is important for policy reasons to know whether it is the schools or the peers that drive the positive effects of attending attractive schools. If it is the schools, then one policy implication is that one should study the successful schools so as to learn from and adapt their approaches in other schools. School effects also point to the importance of recruiting and retaining good teachers and indicate that increasing the resources available to schools for enhancing quality will improve student outcomes. On the other hand, if it is the peers at good schools who are responsible for the observed positive effects, increasing school resources could be a needless use of public resources.

One aim of this paper is to disentangle school effects from peer effects. Using a school choice reform, we argue that we were able to identify peer effects on a group of students who did not change school type. The treatment effect of the school choice reform can be written as:

$$Y_{1i} - Y_{0i} = \delta_i$$

The effect of the intake reform is not constant across students:

$$\delta_{ad} = \beta_{ad} + \mu_{ad} + \rho_{ad} + \sigma_{ad},$$

where a indicates high ability students, d indicates downtown students, β_{ad} is the peer effect, μ_{ad} is the school effect, ρ_{ad} represents the effect of a change in travel distance, and σ_{ad} is the incentivizing effect of the school choice reform. Since we restrict the sample to high-ability downtown students, we assume initially that the intake reform does not affect travel distance or type of school, since these students still attend nearby schools of the same quality. However, we find that high-ability downtown students move systematically *between* downtown high schools as a consequence of reform. This may be an indication that downtown schools are not all of the same type, and that high-ability downtown students actually do experience school effects. Nevertheless, since movement of high-ability downtown students between downtown schools is limited, only a fraction of the identified effect can be attributed to potential school effects, even if there are substantial differences in school quality.

1.3 Universal Childcare, Childcare Quality, Starting Age, and School Performance

The effects of early childhood education have gained increasing interest among social scientists, politicians and especially economists. There is now an emerging consensus that the positive effects of high-quality childcare can be significant for children from disadvantaged backgrounds (Anderson, 2008; Heckman et. al, 2010). A next step for research is to examine whether it is possible to scale up those interventions, which have previously been effective in high-cost programs of limited reach, and achieve the same effects. The literature on targeted programs can also be extended by examining the effects for different groups of children, such as those not from disadvantaged backgrounds or very young children. The main way to answer these questions comes from studying large-scale public childcare programs, regarding which the literature has thus far provided mixed conclusions.[1] Performing randomized experiments on such a large scale in this setting is generally considered unfeasible, so natural experiments are used to identify treatment effects. Baker, Gruber, & Milligan (2008) is an early example of a study on universal childcare programs in the 1990s in Canada. The Province of Quebec introduced a program that greatly increased the level of subsidies for childcare places. As a consequence, childcare attendance increased in Quebec compared to other Canadian provinces. Comparing the measures of health and behavioral outcomes in Quebec with the rest of Canada they find evidence that children in Quebec are worse off on several dimensions.

Havnes & Mogstad (2011) looked at a natural experiment from the Norwegian childcare system. Their study used a 1970s reform that led to a large-scale capacity increase in childcare in only a few years. Since the reform was implemented several decades ago, the authors were able to look at the adult economic outcomes of the children affected by the

[1] “Childcare” is the common UK term employed in this thesis; “day care” is common in other countries.

[2] In both Norway and Denmark, family daycare is mostly a home-based care alternative that normally

reform. Comparing children living in municipalities that had high childcare coverage expansion to those from municipalities with low coverage expansion, the study finds positive effects of universal childcare on adult labor market outcomes.

This article contributes to the literature by studying the effect of a universal childcare expansion on child outcomes. While Havnes & Mogstad (2011) looked at a capacity expansion that occurred in the 1970s, this article uses a more recent large-scale expansion as a source of plausible exogenous variation in childcare attendance. The 2000s have seen a rapid increase in Norway's number of young children in childcare. From 2000 to 2010, the proportion of children aged one and two in childcare increased from 38% to 79%. The increase in capacity can be attributed in part to "The Childcare Agreement" reform of 2003, when several measures were implemented to increase childcare coverage across the country. There had been large variations in the existing coverage for one- and two-year-olds across municipalities. The reform included several measures that led municipalities with low coverage to increase their coverage rates to a greater extent than municipalities that already had higher coverage rates. Pre-reform childcare coverage rates are thus a strong predictor of the level of childcare expansion in the 2000s. The empirical strategy employed in this article relies on comparing the outcomes of children that live in municipalities with low-pre reform coverage (high expansion) to children that live in municipalities with high pre-reform coverage (low expansion) before and after the reform. Differences in changes in test scores could then be attributed to the childcare expansion.

1.3.1 Empirical specification and findings

The empirical specification, using pre-coverage rates, follows Duflo (2000, 2004) in a school setting and Løken et. al (2017) in an eldercare setting. The specification is based on a difference-in-difference strategy (DD). The main difference from a standard DD strategy is that we rely on a pre-reform indicator to measure the intensity of the childcare expansion over time, instead of having a treatment and comparison group of municipalities measured both before and after an expansion. The contrast with a standard DD specification can be illuminated trough a simple two-period formal example. Individual test scores can be written as:

$$Y_{ij}^{post} = \gamma_j + (\rho + \beta)P_j + \mu + \varepsilon_{ij}$$

$$Y_{ij}^{pre} = \gamma_j + \rho P_j + \eta_{ij},$$

where Y_{ij}^{post} is the 5th grade test score for individual child i in municipality j for cohorts born after ($post$) the childcare expansion, Y_{ij}^{pre} is the test score for cohorts of children born before the expansion, γ_j is a time-invariant municipality fixed effect reflecting the fact that children in different municipalities score differently, μ is a municipality-invariant time effect that indicates the common change in test scores from before to after the expansion, ε_{ij} and η_{ij} are error terms reflecting all other factors that can influence test scores, P_j is the coverage rate measured before the expansion, ρ indicates the relationship between test scores and childcare coverage rates before the expansion, and β shows how this relationship changes after the expansion. For simplicity, we keep to the two-period case here, as it is easily extended to the standard regression DD model. Following the notation used by Duflo (2004), the pre-post difference can be written as:

$$\bar{Y}_{ij}^{post} - \bar{Y}_{ij}^{pre} = \beta P_j + \mu + \epsilon_{ij} \qquad (1)$$

Our main interest in this article is to estimate β. Writing the model as in Eq. (1) makes especially clear the assumption upon which the identification relies. $\bar{Y}_{ij}^{post}, \bar{Y}_{ij}^{pre}$ are the average municipality test scores. An estimate of β can be obtained by regressing the change in municipal-level average test scores from before and after the reform on pre-reform coverage rates. The fact that different types of municipalities have high or low coverage rates should not influence estimation, since municipal fixed effects are differenced out. That is, we are comparing changes in test scores within municipalities. Our identification relies on P_j's not being correlated with any factors that remain in the error term ϵ_{ij}. This term reflects all other factors that can influence changes in test scores from before to after the reform. Without a pre-reform indicator of expansion, a common procedure in the literature is to replace P_j with a dummy M_j that indicates whether the actual expansion in a given municipality was large or small. The advantage of our method compared to that approach is the fact that it is easier to accept that ϵ_{ij}, which can be seen as a change in unobserved factors that affect test scores before and after reform, is independent of the pre reform coverage rate, than to accept that the actual change in childcare coverage is independent.

An estimate of the parameter β can be obtained by estimating the following regression:

$$Y_{ijt} = \gamma_j + \mu D_t + \beta(P_j \cdot D_t) + v_{ijt},$$

where D_t is the indicator for post-reform cohorts. It is this last specification that is used to produce our findings. The results do not indicate that the childcare expansion had any average impact for the children exposed to it.

Municipalities are then split into groups according to where they ranked in the distribution of municipality-level proportions of preschool teachers among pedagogical leaders (pedagogical leader are a childcare position type that requires certified education). The group of municipalities with the highest proportion of preschool teachers among pedagogical leaders is called "high-quality" municipalities, while the group with the lowest is called "low-quality" municipalities. Estimating the effect of childcare expansion on high-quality municipalities, we find positive effects on child test scores, while we find negative effects on child test scores in low-quality municipalities. It is important to note that observable inputs to childcare are not randomly distributed across municipalities. Even though we find positive effects of the expansion in municipalities with better-educated staff and negative effects in municipalities with less well-educated staff, other explanations are possible. In particular, we observed that the two groups of municipalities increased coverage for different age groups differently. Positive effects are found in the group of municipalities that expanded access mostly to older children, while negative effects are found among municipalities that largely expanded access to children aged one or two years. Based on these findings, an important starting point for further investigation is therefore to disentangle the role of child age and childcare quality in determining the return on attending childcare. Furthermore, the results point to an important heterogeneity in the effects based on child characteristics. The heterogeneity discovered resonates with some of the previous literature on public childcare programs. The positive effect of Norwegian public childcare for three- to five-year-olds is in line with the findings in Havnes & Mogstad (2011, 2015). Negative effects for younger children are similar to the results in Baker, Gruber, & Milligan (2008). Our results are also consistent with Gupta & Simonsen (2010), who found negative effects of family daycare on boys whose mothers had vocational-track education.[2] We find negative effect in low-quality municipalities on children in low-socioeconomic status families, and no positive effect on this group in high quality municipalities. Our findings suggest that examining how quality, age, and these child characteristics interact is important to be able to evaluate more accurately how public childcare programs affect children's future outcomes.

[2] In both Norway and Denmark, family daycare is mostly a home-based care alternative that normally cares for younger children and is often run by parents. This form of care is public subsidized in Norway.

Large expansions of childcare programs that influence different groups of children are likely to produce heterogeneous effects.

1.4 Households' Responses to Price Changes in Formal Childcare

Nominal childcare prices in Norway fell from 2002 to 2010; in the same period, there was a significant increase in childcare attendance.[3] From 2002 to 2010, childcare attendance for children aged one or two rose from 41% to 79%. Increased capacity, family structure, attitudes, childcare quality, and price are all factors in explaining this growth in childcare attendance. This article seeks to determine how childcare utilization responds to a change in the childcare price. Norway's Cash-for-Care (CFC) reform was enacted on 1stAugust 1998. By 2002, it provided 3,000 NOK (1€ ~ 7.5 NOK in 2002) monthly for each one- or two-year-old child that they did not send to childcare. The size of the benefit corresponded to about 108% of the price of childcare. To analyze the consequences of this price change on formal childcare attendance, we compare differences in childcare attendance rates of eligible children aged one or two to non-eligible children aged three to five before and after reform.

Using the CFC reform, we find that childcare attendance of one- and two-year-olds declined by 14.4 percentage points by 2002, corresponding to a price elasticity of -0.35. From a public policy perspective, it is important to analyze how childcare attendance responds to prices, especially in a regulated market in which the authorities have significant influence in setting prices. However, it must be emphasized that the response to a price change in 2002 might be very different than to a similar price change in 2017, for several reasons. Households at the margin of enrolling a child in childcare in 2002 and 2017 may be different, with their decisions perhaps depending on the income level. There may also be different levels of excess demand. A change in attitude that places more expectations on mothers to re-enter the labor market earlier after birth, which would make price less influential in the enrollment decision, could also be important. These factors counsel caution about

[3] The household surveys used in the analysis show that the average cost of childcare for one child in 2002 where 2707, while it was 2110 in 2010. The 2010 amount adjusted for inflation is 1804.

extrapolating the findings of this study to offer conclusions about how a similar benefit introduced in 2017 would affect childcare attendance.

In addition to advancing the understanding of the response to a price change in childcare on formal childcare participation, we present evidence on alternative modes to public childcare. To understand in appropriate depth how increased childcare attendance affects children in the long run, it is essential to understand what the alternatives are. For example, the effect of increasing childcare attendance may differ if the counterfactual mode of care is informal caregivers as opposed to parental care. The results suggest that parental care is the most important counterfactual care arrangement, since this form of care increased 9.4 percentage points after the 1998 reform. Nannies appear somewhat less important, with an increase of 3.6 percentage points.

In 2002, about 41% of children aged one or two and 84% of children aged three to five attended childcare. Availability, family structure, and childcare quality are arguably constant across children in these two age groups at any given time. The main explanations for differences in attendance rates between these two groups are therefore differences in preferences for childcare for younger and older children and price differences between the two groups. The CFC benefit is the main reason why prices for childcare slots differ for the two groups of children. If there were no difference in parental preferences regarding childcare for younger and older children, it could be argued that the CFC benefit would completely explain the gap in attendance rates.

The difference in price between 2002 and 2010 and the difference in price between children aged one or two and children aged three to five provide two potential ways of calculating the price elasticity of childcare. For example, the inflation adjusted decline in average parental payment from 2002 to 2010 was (2707 – 1804) = 903 NOK. Calculating elasticity with this method gives ((0.79-0.41)/0.41)/((1804-2707)/2707) = -2.78.

Alternatively, one can use the difference in price across age groups in 2002 to calculate the price elasticity for a middle-income household: ((0.82-0.40)/0.40)/(3000/2600) = 0.91. Both numbers could be considered upper bounds on the price elasticity and thus suggest that other factors prominent in the period 2002–2010 have contributed to the rise in childcare attendance for one- and two-year-olds. There has been a great increase in availability, especially following "The Childcare agreement" reform of 2003.

Understanding how childcare attendance responds to childcare prices is important from a public policy perspective. Using easily available estimates to estimate the response may provide deeply misleading assessments of how attendance rates respond to childcare prices. This article uses a 1998 reform that substantially increased the price of childcare for one group of children, while leaving it unchanged for another. This provides us with a causal estimate of the effect of a price change on childcare attendance. At the same time, it allows us to analyze different aspects of household behavior that have not previously been studied in great depth.

1.4.1 Model and previous literature

Identification relies on comparing the differences in childcare attendance rates of eligible children (aged 1–2) to non-eligible children (aged 3–5) before and after reform. The DD model can be specified through the potential outcomes framework. Let Y_{1t} and Y_{0t} be the potential childcare use for an individual with and without a price change in formal childcare:

$$Y_{1i} - Y_{0i} = \delta$$

$$E(Y_{0iat}) = \gamma_a + \lambda_t,$$

where a indexes age group, i indexes the individual child, t indexes time, and δ is the causal effect of the policy. In the absence of any price changes, childcare attendance is determined by a time-invariant age effect (γ_a), and an age-invariant time effect (λ_t). Let D_{at} be a dummy indicating children aged 1–2 after the implementation of the CFC reform. Observed childcare attendance can then be written as:

$$Y_{iat} \ = \gamma_a + \lambda_t + \delta D_{at} + \varepsilon_{ist},$$

where ε_{ist} is an error term that includes other factors that can determine attendance rates.

The 1998 CFC reform has previously been analyzed in studies with a primary focus on its effect on parents' labor force participation. For example, Naz (2004) compared the labor force participation of parents of one- and two-year-olds and parents of three- to five-year-olds before and after reform. This is similar to the identification strategy described above. The main conclusion of that article is that specialization within the household increased following the reform: mothers decreased their labor market participation, while there was little change in their husbands' labor force participation. Using administrative data, Schøne (2004) shows that the effect is somewhat smaller after controlling for macroeconomic factors by employing a difference-in-difference-in-difference strategy.

This study adds to the previous literature studying Norway's CFC reform by focusing on the effects of a price change in childcare on the children involved. Part of the motivation behind this approach is to help explain the results found in studies of the effects of childcare. For this purpose, this article contains an investigation of what the alternative mode of care is for one- and two-year olds. While we use the CFC reform here, these reveal something general about the price sensitivity of parents to childcare prices. The CFC reform constituted a large shock to childcare prices not easily found in other contexts that offers us the ability to its consequences.

References

Abdulkadiroğlu, A., Angrist, J., & Pathak, P. 2014. The elite illusion: achievement effects at Boston and New York exam schools. *Econometrica*, 82 (1), pp.137–196.

Abdulkadiroğlu, A., Angrist, J.D., Dynarski, S.M., Kane, T.J., & Pathak, P.A. 2011. Accountability and flexibility in public schools: evidence from Boston's charters and pilots. *The Quarterly Journal of Economics*, 126 (2) (2011), pp.699–748.

Altonji, J.G., Elder, T.E., & Taber, C.R. 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113 (1), pp.151–184.

Anderson, M.L. 2008. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103 (484), pp.1481–1495.

Angrist, J.D. 2014. The perils of peer effects. *Labour Economics*, 30, pp.98–108.

Angrist, J.D., Cohodes, S.R., Dynarski, S.M., Pathak, P.A., & Walters, C.R. 2016. Stand and deliver: effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, 34 (2), pp.275–318.

Angrist, J.D. & Lang, K. 2004. Does school integration generate peer effects? Evidence from Boston's Metco Program. *American Economic Review*, 94 (5), pp.1613–1634.

Atkinson, A.B., Rainwater, L., & Smeeding, T.M. 1995. *Income distribution in advanced economies: Evidence from the Luxembourg Income Study (LIS)*. Luxembourg Income Study Working Paper Series 120. Luxembourg: Luxembourg Income Study.

Baker, M., Gruber, J., & Milligan, K. 2008. Universal childcare, maternal labor supply, and family well-being. *Journal of Political Economy*, 116 (4), pp.709-745.

Booij, A. S., Leuven, E., & Oosterbeek, H. 2017. Ability peer effects in university: Evidence from a randomized experiment. *The Review of Economic Studies*, 84 (2), pp.547-578.

Bryk, A., Lee, V.E., & Holland, P.B. 1993. *Catholic schools and the common good.* Cambridge, MA: Harvard University Press.

Carrell, S. E., Sacerdote, B. I., and West, J. E. 2013. From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica*, 81 (3), pp.855–882.

Chetty, R., Hendren, N., & Katz, L.F. 2016. The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. *The American Economic Review*, 106 (4), pp.855–902.

Clark, D. 2010. Selective schools and academic achievement. *The BE Journal of Economic*

Analysis & Policy, 10 (1).

Coleman, J., Hoffer, T., & Kilgore, S. 1982. Cognitive outcomes in public and private schools. *Sociology of Education*, 55 (2), pp.65–76.

Dobbie, W. & Fryer, R.G. 2011. Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, 3 (3), pp.158–87.

Dobbie, W. & Fryer, R.G. 2015. The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, 123 (5), pp.985–1037.

Duflo, E. 2001. Schooling and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *The American Economic Review,* 91 (4), pp.795–813.

Duflo, E. 2004. The medium run effects of educational expansion: evidence from a large school construction program in Indonesia. *Journal of Development Economics*, 74 (1), pp.163–197.

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

Evans, W.N. & Schwab, R.M. 1995. Finishing high school and starting college: Do Catholic schools make a difference? *The Quarterly Journal of Economics*, 110 (4), pp.941–974.

Figlio, D.N. & Stone, J.A. 1999. Are private schools really better? *Research in Labor Economics*, 18 (1), pp.115–40.

Gleason, P., Clark, M., Tuttle, C.C., & Dwoyer, E. 2010. *The evaluation of charter school impacts: final report. NCEE 2010-4029.* Washington, DC: National Center for Education Evaluation and Regional Assistance.

Grogger, J., Neal, D., Hanushek, E.A., & Schwab, R.M. 2000. Further evidence on the effects of Catholic secondary schooling [with comments]. *Brookings-Wharton Papers on Urban Affairs*, pp.151–201.

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

Havnes, T. & Mogstad, M. 2011. No child left behind: subsidized child care and children's long-run outcomes. *American Economic Journal: Economic Policy*, 3 (2), pp.97–129.

Havnes, T., & Mogstad, M. 2015. Is universal child care leveling the playing field?. *Journal of Public Economics*, 127, 100-114.

Heckman, J.J., Moon, S.H., Pinto, R., Savelyev, P.A., & Yavitz, A. 2010. Analyzing social experiments as implemented: A reexamination of the evidence from the Highscope Perry Preschool program. *Quantitative Economics* 1 (1), pp.1-46.

Hoxby, C.M., Murarka, S., & Kang, J. 2009. *How New York City's charter schools affect achievement, August 2009 Report.* Second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project.

Hoxby, C.M. & Rockoff, J.E. 2004. *The impact of charter schools on student achievement.* Cambridge, MA: Department of Economics, Harvard University.

Jackson, C.K. Single-sex schools, student achievement, and course selection: evidence from rule-based student assignments in Trinidad and Tobago. 2012. *Journal of Public Economics*, 96 (1), pp.173–187.

Kling, J.R., Liebman, J.B., & Katz, L.F. 2007. Experimental analysis of neighborhood effects. *Econometrica*, 75 (1), pp.83–119.

Kling, J.R., Ludwig, J., & Katz, L.F. 2005. Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment. *The Quarterly Journal of Economics*, 120 (1), pp.87–130.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., & Sanbonmatsu, L. 2013. Long-term neighborhood effects on low-income families: evidence from moving to opportunity. *American Economic Review,* 103 (3), pp.226–231.

Naz, G. 2004. The impact of cash-benefit reform on parents' labour force participation. *Journal of Population Economics*, 17 (2), pp.369–383.

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

The Extent of Bias in Grading

Leroy Andersland[†]

This version: 30 August 2017

Abstract

Do biased perceptions and behaviors affect teachers' assessment of students? To investigate this question, a number of studies use data on two different scores for the same individuals: one non-blind score based on classroom tests assessed by the student's own teacher and one blind test score based on a national exam marked externally and anonymously. In the absence of bias in teachers' assessments, it is argued, there should not be significant differences in the gaps in blind and non-blind scores between different groups. This article present a parsimonious econometric framework that distills out the assumptions necessary to identify group bias in teachers' assessment from such a comparison of blind and non-blind scores. This framework lays the foundation for our empirical analysis, where data from the Norwegian school system are employed to estimate and interpret differences between non-blind and blind assessments. The results suggest that the relationship between the subject ability and non-blind results tends to be different from the relationship between subject ability and blind results. Evidence of this is found both when grades are recorded when teachers grade the same test and when they grade based on different assessments that are meant to test the same skill. The difference between non-blind and blind will therefore be a function of the skill tested. This leads to different estimates of the group bias when holding ability fixed.

Keywords: Discrimination; bias; human capital; test scores

JEL codes: D80; D63; J15; J16; J24

[†] Department of Economics, University of Bergen, 5020 Bergen, Norway; leroy.andersland@econ.uib.no

1 Introduction

Economists and policymakers are keenly interested in the existence and importance of stereotyping and discrimination by schoolteachers. One question receiving particular attention is whether gender-biased perceptions and behaviors affect teachers' evaluation of students. To answer this question, a number of studies compare teachers' average marking of boys and girls in a classroom exam assessed by the student's own teacher (non-blind scores) to the respective means in a nationally set exam marked externally and anonymously (blind scores). This approach was pioneered in Lavy's (2008) study of gender bias in Israel, and subsequently, it has been applied to data from many other countries (see, for example, Lindahl, 2007; Cornwell, Mustard, & Van Parys, 2013; Burgess & Greaves, 2013).[1] These studies report significant differences across groups in blind and non-blind test scores, and interpret these differences as evidence of stereotyping or discrimination by teachers.

The goal of this paper is to assess whether and in what situations systematic differences between non-blind and blind assessment across groups can be interpreted as evidence of stereotyping or discrimination by teachers. We focus on two types of data generating processes of the blind and non-blind scores. The first type occurs when the student's own teacher and an external examiner are marking the *same* test. As in most previous studies, the second is a data-generating process in which the student's own teacher and an external teacher are marking *different* tests that are meant to measure the student's knowledge of the same material. We present a parsimonious econometric framework that shows, for each data-generating process, the assumptions under which one can draw causal inferences about bias in teachers' assessment from a comparison of blind and non-blind test

[1] Differences between non-blind and blind assessment across groups have been used to measure discrimination or stereotypes in several other settings (see, for example, Blank, 1991; Goldin & Rouse, 2009). An alternative approach to measuring discrimination or stereotyping is to randomly assign certain characteristics (e.g., gender) to students' exam scripts (Hanna & Linden, 2009; Sprietsma, 2013) or job applications (Bertrand & Mullainathan, 2004).

scores. This framework lays the groundwork for our empirical analysis, where data from the Norwegian school system is employed to estimate and interpret differences between non-blind and blind assessment of students.

Importantly for our analysis, the Norwegian data offer information on two sets of blind and non-blind scores. One set of scores is generated by assessment of the same test by examiners that do not know the identity of the student and the student's own teacher. The other set of scores comes from assessment on different tests (testing the student's knowledge of the same material) by external examiners and the student's own teacher. As in previous studies, the results show that the scores of boys and girls differ significantly in the non-blind classroom assessments marked by the student's own teacher as compared to the scores in a nationally set exam marked remotely and anonymously by an external examiner. If data from two evaluations of the same test are used, a similar difference appears, though it is not statistically significant. A possible explanation for a potential difference between the two types of data is that females tend to perform better than boys in classroom tests assessed by their own teacher as compared to nationally set exams marked by an external examiner. Another is that female students are better at a potential skill only tested in teacher assessment compared to boys. The result shows that the relationship between subject ability and non-blind grades is different from the relationship between subject ability and blind grades. This is found even when teachers grade the same exam. This leads to different estimates of the group bias when holding ability fixed.

The remainder of the paper proceeds as follows. The next section provides background on the Norwegian school system, discusses how exams are set and assessed, and describes our data. Section 3 presents the econometric framework, laying out the possible sources of differences in blind and non-blind test scores. Section 4 describes and discusses our findings, and the final section offers some concluding remarks.

2 Institutional Background and Data

This analysis employs data comparing exams that are graded externally and anonymous with local teacher evaluations. These records will be referred to as the administrative data. In addition, we have been given access to files from experiments in two different areas comparing the same test graded anonymous and by the students' teacher. These records are called the non-administrative data. This section gives an overview of the education system, focusing on the importance of the tests, how grading is undertaken, and a data and variable description.

2.1 The Norwegian Education System

The Norwegian pre-college education system consists of primary school (level 1-7), lower secondary school (level 8-10), and upper secondary school (level 11-13). Both primary and lower secondary schools are compulsory. The majority of students attend a public institution, and even private institutions are funded and regulated by the Ministry of Education and Research. There are generally no tuition fees.

Norwegian municipalities operate primary and lower secondary schools. At the primary school level, all students are allocated to schools based on fixed school catchment areas within municipalities. With the exception of some religious schools and schools using specialized pedagogic principles, parents are not able to choose the schools to which their children are sent (except by moving neighborhoods). There is a direct link between elementary school attendance and attendance at middle or lower secondary schools (ages 13–16/grades 8–10), in that elementary schools feed directly into lower secondary schools. In many cases, primary and lower secondary schools are also integrated. At the end of middle school, students are evaluated both non-anonymously by their teachers for most subjects taught in school, and in addition anonymously and externally in one to two central exit exams.

At the end of 10th grade, students apply for upper secondary school. The high schools have two main tracks, vocational and academic. They are administered at the county level (above the level of municipalities) and are not mandatory in Norway, although, since the early 1990s, everybody graduating from middle schools is guaranteed a slot in high school.

Admissions procedures differ across counties for upper secondary schools. In most counties, students can freely choose schools, but in others, children are allocated to schools based on well-defined catchment areas, or high school zones. In both regions we focus on, students are free to choose schools within their regions. This means that middle school grades are important for intake to schools and tracks where there is competition.

At the middle school level, the final Gradepoint is based on teacher evaluations of in-school performance, as well as central exams. The Gradepoint summarizes student performance at school, and is used for track and school placement later. Both oral and written performance are assessed in some subjects, and both oral and written exams are given. Our data show that, in the period 2000–2010, students had on average 14.0 teacher-given grades and 1.37 written exam grades and 1.0 oral exam grades in middle school.[2] In middle school, the Gradepoint consists of the average grade times 10, where all topics (exams and in-school assessments) have a grade between 1 and 6. A new Gradepoint is calculated at the end of high school, and is the average grade on all high school exams and subjects times 10. In addition to Gradepoints, points are given according to specific criteria to make up the final measure that determines school and track placement at post-secondary education. It is also common to attach high school certificates showing grades to job applications.

[2] In the same period, students had on average 22.4 teacher-given grades and 6.5 exams (oral and written) at academic track high schools.

2.2 Grading

Grading principles are set by the *Education Act* (*Opplæringslova*). It is stated that teacher course evaluations shall be based on to what degree students have achieved the competence goals, stated by the subject-specific and nationally set learning goals. For most subjects, the final teacher evaluation grade is set based on achieved competence in the late spring each year. Notably, it is specifically stated that student behavior (*orden og oppførsel*) is not to be reflected in grading, and (of course) that student background should not count in grading. Effort is allowed to be included in grading in gymnastics. Teacher course assessment grades are set before the grading of exams. Normally, schools have a local test called *Tentamen* near the end of each semester in middle school. It is an important part of the teacher's final evaluation.

Students do not have an exam in each subject. At the end of middle school, students are drawn to take Norwegian, Math, or English written exams. Students drawn to Norwegian perform two exams, one for each official written language (Bokmål and Nynorsk). The written exam is nationally prepared and corrected by two sensors that are external to the school and who do not know the identity of the student. Students are also drawn to perform an oral exam in any subject, which is administered at the local level. The exams are part of the evaluation of the students' achieved competences in a subject according to the centrally set learning goals. The learning goals have an oral component in some subjects. We focus on teacher assessments in two subjects, Math and written Norwegian, where the oral component does not matter. Thus, teacher assessments and the national exams we use are supposed to test the same skills.

For two regions, Rogaland and Bergen, we have the non-administrative datasets. In the spring of 2015, the school authorities in the municipality of Bergen conducted an experiment on all students at middle schools in Bergen. For the high-stakes *Tentamen* at the end of 10th grade, an additional teacher graded the tests anonymously, in addition to the students' teacher in that topic. All students take the *Tentamen* in Math, English, and Norwegian, but it varies by class in which subject an additional teacher graded the test anonymously. We have information on the gender of the students, as well as whether they are immigrants or not for a part of this sample. The teacher that was to grade the test anonymously was another teacher at the same school. Therefore, it is likely that all teachers knew that this experiment took place.

For Rogaland, we have a similar dataset at the high school level. Here, a student's tests were graded both by teachers at the same school and by an external group of examiners elected by the county-level school authorities. The test is a locally administered end-of-year exam. In contrast to *Tentamen*, the grade on this exam appears as a separate grade on the students' certificates and counts in calculating their Gradepoints. In addition, students' names do not appear on their exams. This is different from centrally administered exams in that the students' local teachers participate in both making and grading the exam.[3] In addition to the external group of examiners that grade the tests anonymously for the blind evaluation, two teachers grade the locally administered exams for the non-blind evaluation, one of which is the student's teacher. The other is a teacher external to the school. The procedure in the first year, 2010, was that 6 schools were drawn to provide 10 exams each (at random) and submitted to the school authorities. Then a group of external examiners were chosen to grade the exams. Half of the tests this year were in Mathematics and half were in Norwegian. In 2012 and 2013, the experiment was followed up and extended to include more schools. In these years, the schools were also randomly drawn. We do not have any observable characteristics of the students for this dataset. Nevertheless, we can still examine the pattern in non-blind and blind scores, and compare it to the results from in the administrative data.

[3] In Math one part of the exam is made at the county level, and one part on the school level. In Norwegian the whole exam is made on the county level.

The experiment in Bergen was performed at the middle school level, while the Rogaland experiment was performed at the high school level. The comparable administrative data is at the same school level. As discussed in this section, there are some differences between scores within the non-blind and blind definitions. Table 1 summarizes institutional details about the grader, number of graders, etc.

[Table 1]

One thing to note from Table 1 is that in the data from Rogaland, the blind grade in the administrative data is the same as the non-blind in the non-administrative dataset. For the administrative data, the score on the local exam is defined as blind since the name does not appear on the test, while the teacher knows the student's identity for course assessments. In the non-administrative data, the same score on the local exam is non-blind since the student's teacher grades the exams, while external examiners give the blind scores.

Since 2012, the standard has been that exams in Norwegian are written on a computer, while exams are written on both paper and digitally in Mathematics. For the experiment conducted in Bergen, the Norwegian tests are written digitally, while the Mathematics tests are written on paper.

Failing a course assessment or an exam (local or external) in middle school, the student will still be able to attend high school, but it may have consequences for the student's options regarding track and school placement. If the student fails a compulsory course assessment or compulsory exam in high school, he or she will not be able to complete the education in that track. Failing the *Tentamen* does not have any direct consequences other than being negative for the course assessment.

2.2.1 Variable Definitions

For the administrative dataset, we are able to match middle and high school grades to register-based files. Students are defined as having a low socio-economic-status (SES) if none of their parents have completed college/university and the father has earnings below the 50th percentile in the income distribution of fathers in the sample. Students are categorized as immigrants if they have one or two parents born in a non-Western country. The register-based files also provide information on the student's gender. The grade files provide information on which school the student attended for each year the grade is registered.

We have split the administrative data into three main samples. The first is a sample of grades given to students at the same schools, years, subjects, and level as in the experiments. The second is a sample of grades from all middle schools/high schools in Bergen/Rogaland from the same years, subjects, and level, while the third contains grades from these areas given in the period 2008–2015 for the same subjects and levels. The grades in the non-administrative data are from 2015 in the Bergen experiment, while grades are from 2010, 2012, and 2013 in the Rogaland experiment.

School administrators supplied data from the experiments directly to us. In the Bergen experiment, we were able to derive grades, school, year, gender, class, subject, and a personal identifier. For the Rogaland experiment, in addition to grades, we have information on the school, subject, and year.

3 Setup

3.1 Notation and Modeling

3.1.1 Data-Generating Processes

The data are from two different data-generating processes.[4] The first is the non-administrative data, which are based on the experiments that assigned the same test in Bergen to be graded by different examiners. Here, we can observe student i at school (or class) s taking only one

[4] The model is explained in terms of the institutional setup for Bergen, since it is here we compare non-administrative and administrative group coefficients. Differences in the setup for Rogaland are presented in the institutional details and data section, and will be discussed in the results section.

test. The grade student i receives from her teacher is Y_i^n (non-blind grading result), whereas the one from the other grader is Y_i^b (blind grading result of the same exam). Therefore, we define the grade difference using non-administrative data as Δ_i^e, where

$$\Delta_i^e = Y_i^n - Y_i^b.$$

The second set of records is the administrative data. For this data-generating process, we can observe student i at school (or class) s now taking two different tests: a blind exam and a teacher assessment at her own school, which is graded by her own teacher. The grade student i receives from her teacher is $\tilde{Y}_i^n$, whereas the grade from the external graders is Y_i^b. Therefore, we define the grade difference using administrative data as Δ_i^o, where

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b.$$

3.1.2 Grades Given by Students' Own Teacher in Non-Administrative Data

Let us assume that Y_{is}^n , the grade given by the teacher in the non-administrative data, can be written as

$$Y_{is}^n = t(X_{is}) + \varrho\theta_{is} + \varepsilon_{is}^n.$$

The function $t(\cdot)$ expresses how the teacher at school (class) s affects student i's grade. We assume that $t(\cdot)$ has the following functional form:

$$t(X_{is}) = t(G_{is}, \kappa_{is}, \bar{\theta}_{is}) = \alpha + \beta G_{is} + \gamma\kappa_{is} + (1 - \varrho)\bar{\theta}_{is}.$$

The function $t(\cdot)$ is the **biased grading function**, or simply, **bias**. It describes how teachers bias grades according to student characteristics. The variable X_{is} is a vector that contains G_{is}, which are some observable characteristics to the researcher and the teacher. κ_{is} represents student behavior in class, and $\bar{\theta}_{is}$ is a compound of other information about the students that the teacher uses to grade. In particular, $\bar{\theta}_{is}$ is, for example, other student abilities/behavior, grades in other subjects, or previous grades. κ_{is} and $\bar{\theta}_{is}$ are not necessarily observable to the researcher. The variable θ_{is} is the true ability being measured. The parameter ϱ reflects the relationship (mapping) of that ability to the score given by the teacher. ε_{is}^{n} is an error. The parameter α captures grade inflation, β captures discrimination in favor of groups of students with observable characteristics G, and γ and $(1-\varrho)$ capture the effect of components that are unobservable to us but that are used by the teacher when grading exams.

There are two components of ε_{is}^{n}. The first one, ξ_{is}^{n}, is specific to the grader when assigning a grade to student i. The second one is a component reflecting the student's idiosyncrasy, ϵ_{is}, which is not related to the grader. For example, ϵ_{is} may be any deviation (luck, not feeling well on the day of the internal exam, etc.) that makes the student’s grade not reflect exactly his or her level of ability θ_{is}. Thus,

$$\varepsilon_{is}^{n} = \xi_{is}^{n} + \epsilon_{is}.$$

For those reasons, we rewrite the previous equation for Y_{is}^{n} as

$$Y_{is}^{n} = \alpha + \beta G_{is} + \gamma \kappa_{is} + (1-\varrho)\bar{\theta}_{is} + \varrho\theta_{is} + \xi_{is}^{n} + \epsilon_{is} \quad (1).$$

3.1.3 Grades Given by Students’ Own Teachers in Administrative Data

Let us assume that $\tilde{Y}_{is}^{n}$, the grade given by the teacher in the administrative data, can be written as

$$\tilde{Y}_{is}^{n} = t(X_{is}) + \tilde{t}(X_{is}) + \rho\theta_{is} + (1-\rho)\tilde{\theta}_{is} + \tilde{\varepsilon}_{is}^{n}.$$

$t(X_{is})$ is the biased grading function in the administrative data. The parameter ρ reflects the relationship of ability θ_{is} to the score given by the teacher. $\tilde{\theta}_{is}$ measures a compound of other abilities that are captured by the teacher grade in administrative data. Note that an important difference here from the non-administrative data is that we do not include this term in the biased grading function. This is because, in the administrative data, the two different tests can actually measure different subject skills. The function $\tilde{t}(X_{is})$ explains why some students perform relatively better under in-class tests graded by the teacher. Finally, $\tilde{\varepsilon}_{is}^{n}$ is some variation, containing, for example, an error that is due to the grader, $\tilde{\xi}_{is}^{n}$, and another coming from the student, $\tilde{\epsilon}_{is}$, as he or she may have different performance at another time due to various causes. That is,

$$\tilde{\varepsilon}^{n}_{is} = \tilde{\xi}^{n}_{is} + \tilde{\epsilon}_{is}.$$

Let $t(X_{is})$ be

$$t(X_{is}) = t(G_{is}, \kappa_{is}) = \alpha + \beta G_{is} + \gamma \kappa_{is}.$$

We write $\tilde{t}(X_{is})$ as

$$\tilde{t}(X_{is}) = \tilde{\alpha} + \tilde{\beta} G_{is} + \tilde{\gamma} \kappa_{is}.$$

$\tilde{Y}^{n}_{is}$ is then

$$\tilde{Y}^{n}_{is} = \alpha + \beta G_{is} + \gamma \kappa_{is} + \tilde{\alpha} + \tilde{\beta} G_{is} + \tilde{\gamma} \kappa_{is} + (1-\rho)\tilde{\theta}_{is} + \rho \theta_{is} + \tilde{\xi}^{n}_{is} + \tilde{\epsilon}_{is} \quad (2).$$

3.1.4 Grades Given by External Reviewers

The grade given on the exam from the external grader is Y^{b}_{is}.

$$Y^{b}_{is} = \theta_{is} + \varepsilon^{b}_{is}.$$

We then write

$$\varepsilon^{b}_{is} = \xi^{b}_{is} + \epsilon_{is},$$

where ξ^{b}_{is} is the measurement error that is specific to the external evaluator when assigning a grade to student i and ϵ_{is} is the same term that explains deviations between grades and skills that appeared as a component of ε^{n}_{is}. We therefore rewrite the equation for Y^{b}_{is} as

$$Y^{b}_{is} = \theta_{is} + \xi^{b}_{is} + \epsilon_{is} \quad (3).$$

3.2 Parameters of Interest

The biased grading function $t(\cdot)$ is unknown and is the main object of interest. We want to learn how teachers distort grades. For example, as in Lavy (2008), do teachers favor girls? Or is it another reason for this difference, as suggested in Hinnerich, Hoglin, and Johanneson (2011).

Identification of the parameters α, β, and γ is not feasible without imposing some untestable assumptions. For example, we do not observe κ_{is}, $\tilde{\theta}_{is}$, or $\bar{\theta}_{is}$, which may be arbitrarily correlated with G. However, the relevance for explaining outcome differences between groups of separating out the effect of those variables is not clear, as all can have an effect on future outcomes. In what follows, we show what can be identified from the non-administrative data we have available. We also show what under different assumptions can be identified by administrative data. The main threat to identifying relevant bias in the administrative data would be the function $\tilde{t}(X_{is})$ and the difference between $\tilde{\theta}_{is}$ and $\bar{\theta}_{is}$. If some students perform better at in-class exams, or if the teacher assessments and national exams actually tests different skills, then this should not be characterized as bias.

3.3 Identification Using the Non-Administrative Data

3.3.1 Identification Using Non-Administrative Data: $\varrho = 1$

We have that the variable that measures differences in grades, Δ_i^e, in the non-administrative data can be written as

$$\Delta_{is}^e = Y_{is}^n - Y_{is}^b = t(X_{is}) + \tau_{is} \quad (4),$$

where

$$\xi_{is}^n - \xi_{is}^b = \tau_{is}$$

captures differences in error terms coming from the fact that grades are given by two different people (teacher, ξ_{is}^n, and external reviewer, ξ_{is}^b) for the same exam. We assume that the error τ_{is} is idiosyncratic and not related to any of the other variables on the right-hand side. The differences in grades are equal to $t(\cdot)$ plus the unobserved component τ_{is}:

$$\Delta\,_{is}^{e} = \alpha + \beta G_{is} + \gamma \kappa_{is} + \tau_{is}.$$

Identification of the parameters α, β, and γ is not possible without further assumptions because G and κ_{is} are arbitrarily correlated. In this case, the unobserved component τ_{is} is uncorrelated to the function $t(X_{is})$ and is not the reason the structural parameters of the biased grading function are not identified. Even though we cannot identify α and β, we can identify the parameters of the regression of Δ_{is}^{e} on G:[5]

$$\bar{\beta} = \frac{Cov(\Delta^{e}, G)}{Var(G)} = \beta + \gamma \frac{Cov(\kappa, G)}{Var(G)} \qquad (5)$$

The parameter $\bar{\beta}$ can be interpreted as the total effect of a given characteristic G on the differences in grades. For example, suppose that teachers do not favor girls ($\beta = 0$), but that girls are typically better-behaved in class than boys and that teachers reward girls for their behavior. Thus, $Cov(\kappa, G)$ and γ are both positive. In that case, $\bar{\beta}$ is positive even though β equals zero. Nevertheless, given that G and κ are correlated, any intervention that tries to minimize bias in grading will necessarily be a policy whose overall effect will be measured in terms of $\bar{\beta}$, not β. The intercept $\bar{\alpha}$ can be written as

$$\bar{\alpha} = E(\Delta^{e}) - \bar{\beta}E(G) = \alpha - (\beta + \gamma \frac{Cov(\kappa, G)}{Var(G)})E(G) \qquad (6).$$

Again, the mean bias, α, is not identifiable, but the parameter that will be used to measure the effectiveness of bias on outcomes is not α, but $\bar{\alpha}$.

3.3.2 Identification Using Non-Administrative Data: $\varrho \neq 1$

The difference is then equal to the function of interest, $t(\cdot)$, a function of the skills being measured. Moreover, the unobserved component τ is:

$$\Delta_{is}^{e} = \alpha + \beta G_{is} + \gamma \kappa_{is} + (\varrho - 1)\theta_{is} + (1 - \varrho)\bar{\theta}_{is} + \tau_{is} \qquad (7).$$

Even though we cannot identify α and β, we can identify the parameters estimated by a regression of Δ_{e} on G:

[5] When G is a vector, the usual matrix notation has to be employed. We present the simple regression algebra just to facilitate the exposition of the argument.

$$\bar{\beta} = \frac{Cov(\Delta^e, G)}{Var(G)} = \frac{Cov\big(\alpha + \beta G_{is} + \gamma\kappa_{is} + (\varrho - 1)\theta_{is} + (1 - \varrho)\overline{\theta}_{is} + \tau_{is}, G\big)}{Var(G)}$$

$$\bar{\beta} = \frac{Cov(\Delta^e, G)}{Var(G)} = \beta + \gamma\frac{Cov(\kappa, G)}{Var(G)} + (\varrho - 1)\frac{Cov(\theta, G)}{Var(G)} + (1 - \varrho)\frac{Cov\big(\overline{\theta}, G\big)}{Var(G)} \quad (8).$$

The parameter $\bar{\beta}$ will then consist of gender bias, differences in behavior correlated with gender, and a function of how gender is correlated with the different skills and information that teachers use in setting grades. Importantly, note that, in this case, it is not obvious that $\bar{\beta}$ is the only parameter of interest for evaluating how the total amount of bias affects student outcomes. In particular, it is interesting to know, for a given ability in a subject, the total amount of bias one group receives compared to another. The alternative parameter of interest would be:

$$\bar{\bar{\beta}} = \frac{Cov(\Delta^e, G|\theta)}{Var(G)} = \frac{Cov\big(\alpha + \beta G_{is} + \gamma\kappa_{is} + (\varrho - 1)\theta_{is} + (1 - \varrho)\overline{\theta}_{is} + \tau_{is}, G|\theta\big)}{Var(G)} \quad (9).$$

One way of obtaining an estimate of this would be to insert θ into the right-hand side of Equation (7), using Y^b_{is}:

$$Y^b_{is} = \theta_{is} + \xi^b_{is} + \epsilon_{is}$$

$$\theta_{is} = Y^b_{is} - \xi^b_{is} - \epsilon_{is}$$

Inserting into Equation (7):

$$\Delta^e_{is} = \alpha + \beta G_{is} + \gamma\kappa_{is} + (\varrho - 1)Y^b_{is} + (1 - \varrho)\overline{\theta}_{is} + \xi^n_{is} - \varrho\xi^b_{is} - (\varrho - 1)\epsilon_{is} \quad (10).$$

Because the errors in $-\varrho\xi^b_{is} - (\varrho - 1)\epsilon_{is}$ are correlated with Y^b_{is}, a regression of Δ^e_{is} on G_{is} and Y^b_{is} would not yield the parameter of interest:

$$\ddot{\beta} = \frac{Cov\left(\Delta^e, G|Y^b\right)}{Var(G)} \neq \bar{\bar{\beta}} = \frac{Cov(\Delta^e, G|\theta)}{Var(G)} \quad (11).$$

A solution to this problem is to obtain an unbiased estimate of $(\varrho - 1)$ and fix this parameter in the estimation of Equation (7).

3.4 Using Administrative Data

For the administrative setting, differences in grades can now also be explained by differences in test-type specific performance and differences in the skills that the assessments measure. The grade difference can now be written as:

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b$$

$$= t(X_{is}) + \tilde{t}(X_{is}) + (\rho - 1)\theta_{is} + (1 - \rho)\tilde{\theta}_{is} + \tilde{\varepsilon}_{is}^n - \varepsilon_{is}^b$$

$$= \alpha + \tilde{\alpha} + \left(\beta + \tilde{\beta}\right)G_{is} + (\gamma + \tilde{\gamma})\kappa_{is} + (\rho - 1)\theta_{is} + (1 - \rho)\tilde{\theta}_{is} + \tilde{\xi}_{is}^n - \xi_{is}^b + \tilde{\epsilon}_{is} - \epsilon_{is}$$

In this case it is important to notice that:

- Although $t(X_{is})$ and $\tilde{t}(X_{is})$ are functions of observable (G) and unobservable κ, they have different interpretations. So, a general function $g(X_{is}) = t(X_{is}) + \tilde{t}(X_{is})$ is not measuring biased grading, but the biased grading effect over the fact that some groups of students (e.g., females) perform relatively better under in-class exams than under external exams. Therefore, we cannot necessarily claim that $g(X_{is})$ is biased grading.
- If ρ is different from 1, the grade difference is a function of the competence level of the skill being evaluated. The closer to 1, the smaller the effect of subject-specific ability on the grade difference. Thus, for $\rho < 1$, and as in the non-administrative setting, differences in grades Δ_i^o will depend directly on skills being measured. In contrast to the non-administrative setting, the reason for $\rho < 1$ is not only different grading practices between external and internal teachers, but could also be due to tests measuring different subject skills.
- Unlike in the non-administrative setting, $\tilde{\epsilon}_{is} \neq \epsilon_{is}$, as these two objects come from different exams and luck or feeling ill on an exam day may differ across days.

In what follows, we impose some assumptions that allow us to identify parameters related to the biased grading function using administrative data.

3.4.1 Identification Using Administrative Data: $\boldsymbol{\rho = 1}$

We have that

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b = g(X_{is}) + \tilde{\tau}_{is} \quad (12)$$

$$= \alpha + \tilde{\alpha} + \left(\beta + \tilde{\beta}\right)G_{is} + (\gamma + \tilde{\gamma})\kappa_{is} + \tilde{\tau}_{is},$$

where

$$\tilde{\tau}_{is} = \tilde{\xi}_{is}^{n} - \xi_{is}^{b} + \tilde{\epsilon}_{is} - \epsilon_{is}.$$

As with non-administrative data, we assume that $\tilde{\tau}_{is}$ is idiosyncratic, and that $\tilde{\tau}_{is}$ and X are independent. Thus, one can identify the coefficients of a regression of Δ_i^o on G, exactly as in Equations (5) and (6). The key difference here is that the interpretation of these coefficients would be different, since $\tilde{t}(\cdot)$ is not null. Specifically, they will reflect both bias and differences coming from different test types. Note that both types may explain outcome differences between groups. However, this combined effect could rather be described as the effect of grading schemes rather than bias.

3.4.2 Identification Using Administrative Data: $\boldsymbol{\rho = 1}$ *and* $\boldsymbol{\tilde{t}(\cdot) = 0}$

We have that

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b = t(X_{is}) + \tilde{\tau}_{is} \qquad (13).$$

In this case, the parameters of equations Equations (5) and (6) could be identified.

3.4.3 Identification Using Administrative Data: $\boldsymbol{\rho \neq 1}$ *and* $\boldsymbol{\tilde{t}(\cdot) = 0}$

We have that

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b = \alpha + \beta G_{is} + \gamma\kappa_{is} + (\rho - 1)\theta_{is} + (1 - \rho)\widetilde{\theta}_{is} + \tilde{\xi}_{is}^{n} - \xi_{is}^{b} + \tilde{\epsilon}_{is} - \epsilon_{is} \quad (14).$$

Even though we cannot identify α and β, we can identify the parameters of the regression of Δ_i^o on G:

$$\bar{\beta}' = \frac{Cov(\Delta^o, G)}{Var(G)} = \beta' + \gamma' \frac{Cov(\kappa, G)}{Var(G)} + (\rho - 1) \frac{Cov(\theta, G)}{Var(G)} + (1 - \rho) \frac{Cov(\widetilde{\theta}, G)}{Var(G)} \quad (15)$$

The parameter $\bar{\beta}'$ will then consist of gender bias, differences in behavior correlated with gender, and a function of how gender is correlated with the different skills and information that teachers use in setting grades. Again, it is not obvious that $\bar{\beta}'$ is the only parameter of interest for evaluating how the total amount of bias affects student outcomes. In particular, it is interesting to know, for a given ability in a subject, the total amount of bias one group receives compared to another. An alternative parameter of interest would be:

$$\bar{\bar{\beta}}' = \frac{Cov(\Delta^o, G|\theta)}{Var(G)} \quad (16)$$

A way to estimate total amount of bias conditional on subject-specific ability is to use the blind score:

$$Y_{is}^b = \theta_{is} + \xi_{is}^b + \epsilon_{is}$$

$$\theta_{is} = Y_{is}^b - \xi_{is}^b - \epsilon_{is}$$

Inserting into Equation (14):

$$\Delta_{is}^o = \alpha + \beta G_{is} + \gamma \kappa_{is} + (\rho - 1) Y_{is}^b + (1 - \rho) \widetilde{\theta}_{is} + \tilde{\xi}_{is}^n + \tilde{\epsilon}_{is} - \rho(\xi_{is}^b - \epsilon_{is}) \quad (17)$$

Because the errors in $-\rho(\xi_{is}^b - \epsilon_{is})$ are correlated with Y_{is}^b, a regression of Δ_{is}^o on G_{is} and Y_{is}^b would not yield the parameter of interest:

$$\ddot{\beta}' = \frac{Cov\left(\Delta^e, G|Y^b\right)}{Var(G)} \neq \bar{\bar{\beta}}' = \frac{Cov(\Delta^e, G|\theta)}{Var(G)} \quad (18).$$

A solution to this problem is to obtain an unbiased estimate of $(\rho - 1)$ and fix this parameter in the estimation of Equation (17).

3.5 Comparing Non-Administrative with Administrative Data

Under certain assumptions, administrative data may not be useful for testing for the existence of biased grading. A potential reason for that has to do with the fact that blindly and non-blindly graded exams may differ because these are two different tests. Thus, it is likely that the abilities being measured may be different ($\tilde{\theta} \neq \bar{\theta}$), or that the systematic reaction to the exam may be different ($\tilde{t}(X) \neq 0$). These factors are the main potential reasons resulting from administrative data but do not necessarily identify the same objects as results from non-administrative data. The next section will provide evidence on the difference in estimates produced when using data based on the same exam and data based on different assessments meant to test the same skills.

4 Results

4.1 Descriptive Statistics

Table 2 presents summary statistics of blind and non-blind grades for both non-administrative and administrative datasets. Grades are reported for Math and Norwegian, separately. We also report proportions of students by gender, immigration status, and SES.

[Table 2]

Each column presents the different samples used in the analysis. Column (1) shows summary statistics for the Bergen experiment, while Column (2) shows statistics for middle school grades for the same schools, year, subjects, and level in the administrative data. Columns (3) and (4) show administrative middle school grades for all students in Bergen the same year, subject, and level, and in the period 2008–2015, respectively. In Column (5), statistics from the Rogaland experiment are reported, while Columns (6), (7), and (8) report statistics from administrative samples that include the same schools, years, subjects, and levels as in the experiment; all schools in Rogaland the same years, subjects, and levels; and these grades in Rogaland recorded in the period 2008–2015, respectively.

The total number of observations from the experiment in Bergen is 99. Most students take the *Tentamen* in Norwegian, Math, and English, but only one test for each student was selected for re-grading. For the administrative sample, all students are drawn to perform a national exam in Norwegian, Math, or English. The number of observations is fairly similar to that in the experiment, 105, which is reasonable given the similar system of all students being exposed to anonymous grading in one subject. There are relatively more recordings in Norwegian in the Bergen experiment. For our estimates to be unaffected by the proportion of exams in a particular subject, inverse proportion subject weights are used in the empirical specifications. In the Rogaland experiment, the experiment was carried out by drawing a sample of exams from each school. There are thus more observations for the same schools from the administrative data than in the experiment.

The averages of Math grades are lower than the averages of Norwegian grades, and average blind grades are lower than average non-blind grades. Standard deviations are lower in Norwegian than in Math, but there is not any pattern in the differences in standard deviations between blind and non-blind grades. This does not suggest a leniency bias or centrality bias (Landy & Farr, 1980; Prendergast, 1999). However, if non-blind includes more or different attributes than blind, a centrality bias based on subject-specific ability might not appear as lower standard deviations in non-blind. This is because the different attributes included in non-blind may lead to additional variation in this variable.

Table 3 reports summary statistics of the grade difference between the non-blind and blind grades. Depending on the type of data being used, the grade difference could be a sum of several terms and does not necessarily reflect only the teachers' biased grading, or bias. In administrative data, as discussed in the previous section, differences in grades could be because of teachers' biased grading, that students perform better at one type of exam, or that non-blind and blind grades relate differently to the subject-specific skill. In the non-administrative data, in addition to noise, differences in grades are a sum of teachers' biased grading and that non-blind and blind map differently onto the subject-specific skill. Table 3 presents summary statistics of the grade difference both when aggregating subjects and by subject. Weighted delta (grade difference) is computed by using inverse proportion subject weights. A standardized measure of the difference is constructed by dividing by the standard deviation of the blind exam grade.

[Table 3]

In every subject, the average grade difference is positive. In the non-administrative data from Bergen, the difference is 0.17 standard deviations (SD) of blind exam, whereas in the corresponding administrative data, it is almost three times larger (0.42 SD). In contrast, the non-administrative differences in Rogaland are about twice as large as the difference in the corresponding administrative data. Average differences are smallest in Math, both in absolute terms and relative to variation in blind grades. According to our model, there are several possible explanations for this. The parameters of the biased grading function, $t(X)$, may be different in different subjects. This could, for example, be because there are different types of teachers. Disparities in grade differences across subjects could be explained by the fact that there are differences in student performance across test types in the two subjects. For example, students perform relatively better at in-class exams compared to external exams in Norwegian, compared to Math. Lastly, the students' teachers could weigh skills the students are better at more than external teachers, or in addition, for the administrative data, students are better at the subject skills tested in the non-blind test that are not tested in the blind test. In our model, this would mean that ρ and ϱ differ across subjects, or $\tilde{\theta}$ and $\bar{\theta}$ differ across subjects. A general pattern to notice is that examiners external to the school seem to lead to lower grading. The grade difference is higher in Bergen administrative than non-administrative, while the reverse is true for Rogaland. At the same time, the external graders to the school are grading the blind in the administrative data for Bergen, and in the non-administrative for Rogaland.

4.2 Comparing Estimates of Bias in Administrative and Non-Administrative Data

Table 4 focuses on the non-administrative data from Bergen and compares them to the administrative data from Bergen. The Bergen experiment is particularly interesting because a variable for the observable characteristic *gender* is available, which is used to estimate a coefficient that can be compared to the coefficient obtained from the administrative data from the same year, schools, level, and subject.

[Table 4]

The table presents results from regressions with grade difference as the dependent variable. According to our model, the parameters shown in Equations (5) and (6) would be the correct expressions for the population regression coefficients on group dummies under the assumption that the non-blind and blind relationship to subject skill is the same ($\rho = 1$ and $\varrho = 1$). In addition, for the administrative data, students do not perform differently under different types of tests ($\tilde{t}(X) = 0$).

First, Column (1) shows the results including only subject dummies on the right hand side. Since within transformation on the all of the binary variables used in Table 4 have been performed, the intercept reflects the weighted average grade difference. Adding an indicator variable for gender, Column (2) shows that the gender coefficient is close to 0, with a standard deviation of 0.098. Column (3) displays results when school-interacted fixed effects are included. This increases the gender coefficient to 0.12, though it stays statistically insignificant. Columns (4), (5), and (6) show the same specifications performed on a sample of students from the same schools, years, level, and subjects, using administrative data. The weighted average grade difference is much larger in the administrative data, which may be due to the blind graders being external to the schools. Alternatively, students may perform better on in-class exams, or better at the skills tested by in-class exams, but this is not what was suggested by the data from Rogaland shown in the descriptive statistics of grade differences. An explanation for the higher standard errors in the results in the administrative data is that the variation in student performance across different tests is included as unexplained variation. Results in Columns (1), (2), and (3) and (4), (5), and (6) show that we are not able to reject the null hypothesis of no gender bias in either the administrative or non-administrative data, respectively. The evidence does not suggest that the explanation for the positive gender coefficient in administrative data is that females perform better at in-class tests than external exams compared to males, or that females are better at the skills tested by the in-class tests. Three points are important to note about the non-administrative data for Bergen. First, the low sample size makes it difficult to make any precise statements on the size of gender bias. Thus, the true gender coefficient derived from the comparable sample in the administrative and non-administrative may actually be different. Second, the setup for the Bergen trial make it possible that all teachers knew about the fact that the grading where to be audited. This is different from normal grading of students. Lastly, only two schools were available, and grading in these schools can be different from a representative sample of schools.

In Column (7), we include all students in Bergen. The intercept is relatively similar as in results from the two experiment schools, suggesting similar grading in these schools and the rest of Bergen. Column (8) shows a significant coefficient at the 5% significance level of 0.086, which is close to the estimate with school-interacted fixed effects in the non-administrative data. Including school-interacted fixed effects in Column (9) increases the coefficient to 0.098.

According to our model that describes the content of blind and non-blind grades, there could be several reasons for finding non-blind-blind grade differences that are different across groups in the administrative data. Teacher-biased grading, different performance across test types, and two tests measuring different skills can all be potential explanations. Our results do not suggest that different performance across test types, or that the two test measure different skills, explain the findings for the gender coefficient in the administrative data. However, data limitations make us carful to conclude about the existence of gender bias only relying on the non-administrative sample from Bergen.

4.3 The Relationship between Non-Blind, Blind, and Subject Ability

It is important to determine the relationship between non-blind and subject ability, and blind and subject ability. If these relationships are unequal, it has consequences for the interpretation of grade differences. The slope parameters estimated in Table 4 would then be more correctly described by Equations (8) and (15). For example, grade differences between groups could arise just because the two groups are at different ability levels in the subject (Burgess & Greaves, 2013). There could be several reasons for why the relationship to subject ability differs between the two grades. Tests could measure different skills, or graders are looking at different skills when grading. It is also possible that teachers that know the student avoid giving the student a failing grade, while, for an external grader, it is easier to fail a student.

One approach to evaluate this relationship is to rearrange the variables in our model by inserting Y_{is}^{b} for θ_{is}, as shown in Equations (10) and (17). Blind scores are measuring ability with an error. Therefore, Y_{is}^{b} is correlated with the unexplained part in this equation. Because of this, a simple regression of the grade difference on blind score does not reveal ρ or ϱ, but with a classical measurement error in blind yields a negatively biased estimate of these parameters.

There are two main ways to investigate the importance of measurement error in blind score. First, one could use lagged blind scores as an instrument. The main problem with this procedure is that teachers and the student generally have information on previous and other exam grades the student receives. It is therefore possible that lagged blind grades have a separate impact on non-blind grades. In our model, this is reflected in the terms $\bar{\theta}$ in the non-administrative data and $\tilde{\theta}$ in the administrative data. The other method is to use a regression of the grade difference on grouped average blind score (Deaton, 1985). In our case, the natural way to group students is by school. Table 5 shows these estimations for data from Bergen.

[Table 5]

The table shows results from regressions of the grade difference on individual blind score and school blind score for administrative data in Columns (1)–(4), and for non-administrative data in Columns (5)–(8). Recordings from more years, 2008–2015, are included in the administrative sample to increase precision. The difference between the specification used in Columns (1) and (2) is that school-interacted fixed effects are included in Column (2). This lowers the coefficient on blind scores from –0.25 to –0.22 and suggests that the negative relationship between the grade difference and blind score is partly explained by school factors. The difference between the specifications in Columns (3) and (4) is that blind score-interacted fixed effects are included in Column (4). Column (3) shows a negative relationship of –0.35 between the grade difference and school average blind.[6] This relationship cannot be explained by classical measurement errors in blind score since this is a precise estimate of the school-level ability. There could, however, be other school-level

[6] The regression is performed at the individual level, while school blind is the school average blind for the subject the individual recording is measured in. This specification simply allows for appropriate weighting by school size, while at the same time including subject weights in the regression. School averages are calculated without own recording.

factors that contribute to the negative relationship between group differences and school blind. For example, schools with higher average blind scores are less lenient in non-blind grading. Including blind score-interacted fixed effects reduces the size of the school blind coefficient by 0.21, suggesting that a substantial portion of the negative coefficient is not due to school-level factors. These results are in line with the explanation that a large part of the negative coefficient on blind score is due to non-blind and blind grades mapping differently onto subject ability in the administrative data.

A reason for the negative relationship between the grade difference and school blind could be that non-blind and blind tests in the administrative data actually measure different subject skills. Even though Table 4 did not indicate it, this could lead to the gender coefficient reflecting that females perform better at the subject skills tested in non-blind, but not in blind. Columns (5) and (6) show that the coefficient is only –0.11 using the non-administrative data, indicating that the non-blind and blind grade is more likely measuring the same ability. Still, it also suggests that teachers that know the student, and teachers that do not know the student, grade differently even though they grade the same test. Possible explanations are that the student's teacher knows the identity of the student, the student's class behavior, previous grades, and grades in other subjects. Columns (7) and (8) provide negative, but much less precise, estimates of the relationship between grade difference and school blind. Due to large standard errors, the difference between Column (7) and (8) tells us little, but since both non-blind and blind graders are internal to the school, school-level influences are less likely to be responsible for the negative relationship.

[Table 6]

Table 6 provides results for the separate experiment performed in Rogaland and a comparable administrative dataset. Individual observable characteristics are not available for the non-administrative data, but there are more individual recordings, and recordings from more schools than from the Bergen experiment. Columns (1) and (2) show a negative relationship between grade differences and blind scores of –0.27 and –0.28, respectively. Interestingly, the coefficient increases when including school-interacted fixed effects, suggesting that school-level factors do not contribute to a negative relationship between the grade difference and blind. Column (3) reveals a negative and statistically significant coefficient on school blind of –0.17, which disappears when including blind-interacted fixed effects in Column (4). Note that the student's teacher also grades the locally administered exam, which is used as a blind score for the administrative sample from Rogaland. There is, however, another teacher that also grades the exam, who has less prior information on the student and is external to the school.

As discussed, the negative relationship between the grade difference and blind, indicated by the results provided in Columns (1)–(4), may indicate that non-blind and blind measure different skills. Columns (5) and (6) explore the relationship using the non-administrative data. Column (5) shows a coefficient of –0.17, while the estimate in Column (6) is –0.18. Again, these results do not suggest that schools with higher blind scores are less lenient for the sample of schools from Rogaland. Regressing the grade difference on school blind provides a negative and statistically significant coefficient at the 5% significance level of –0.13 in Column (7). This coefficient increases to 0.05 when including blind-interacted fixed effects. The difference between Columns (7) and (8) is 0.17, which is identical to the coefficient in Column (5).

The results provided in Tables 5 and 6 make it possible to determine the size of 1-ρ and 1-ϱ. Generally, we find a negative relationship when regressing the grade difference on blind. This negative relationship is somewhat smaller in the non-administrative datasets. We also find a similar negative relationship when regressing grade difference on school blind, something that is not explained by measurement error in blind. For the schools from

Rogaland, we do not see any signs that schools with better students are less lenient. Because school-level factors seem to be less of a concern in the Rogaland sample, we use the indicated impact of measurement error in the sample from Rogaland to determine the parameters in Bergen. Using the point estimates of coefficients, the impact of measurement error in blind in administrative data is –0.10 in administrative and –0.04 in non-administrative. Based on results with school-interacted fixed effects from Bergen, this indicates a $1-\rho$ of –0.12 and $1-\varrho$ of –0.05.

Table 5 and 6 provide estimates for $1-\varrho$ from both Bergen and Rogaland. For the non-administrative data from Rogaland, the estimate is larger in magnitude. A possible explanation for this is that the blind evaluator is external to the school in this experiment, leading to a different mapping of grades. In addition, exams were randomly drawn from schools chosen by county level administrative personnel. These features correspond more to the field experiment conduced in Hinnerich et al. (2011), and suggest that the gender bias holding ability fixed is different from the gender bias estimated in that paper.

4.4 Alternative Parameter of Interest

The previous section examined the relationship between non-blind, blind, and subject ability, and found convincing evidence that $1-\rho<0$ and $1-\varrho<0$. This changes the interpretation of the coefficient from a regression of grade difference on the group dummy to now also including a term that reflects the ability difference between groups. In our model, the coefficient is characterized by Equation (8) for non-administrative data and Equation (15) for administrative data. Group bias will now arise if the groups have different subject abilities.

An alternative parameter of interest reflects group bias holding ability constant. For example, this parameter describes the amount of bias a female can expect to get compared to a male of equal ability. In the terms of the model, this parameter is defined in Equations (9) and (16). A way to retrieve an estimate of this parameter is to add blind score as a right-side variable, as shown in Equations (10) and (17). However, this is not feasible using ordinary least squared regression since the model is unidentified, because the blind score is correlated with the unexplained part. Therefore, we use the fact that the last section provided credible estimates of $1-\rho$ and $1-\varrho$, and estimate parameters using constrained least squared estimation, fixing the coefficient of Y_{is}^{b} to the specific values.

[Table 7]

Columns (1)–(3) show results from the non-administrative sample, Columns (4)–(6) for the same schools using the administrative sample, and Columns (7)–(9) all recordings from Bergen in 2015. Column (1) shows the result from an OLS regression of the grade difference on a gender coefficient, shown earlier in Table 4. Column (2) shows results from a constrained least squared estimation, where, in addition to having the gender dummy on the right side, the blind score is included as described in Equation (10). The results confirm that fixing the coefficient on blind to be 0 with this specification gives the same results as an OLS regression of the grade difference on the gender dummy shown in Column (1). Column (3) displays results when fixing $1-\varrho$ to -0.05. The gender coefficient increases to 0.15, but is still not statistically significant.[7] Columns (4)–(6) repeat this procedure for administrative data from the same schools. Column (6) uses the estimate of $1-\rho$ obtained for the administrative data of -0.12. Also here, the gender coefficient increases but is still not significant. In Columns (7)–(9) the gender coefficient is significant for all specifications, and the point estimate of the gender coefficient is very similar to that obtained from the non-administrative data.

[7] Note that this procedure does not account for uncertainty regarding the size of the fixed parameters.

4.5 Other Observable Characteristics

This analysis has compared estimates of gender bias using two different data-generating processes—one where the student's teacher and a teacher that does not know the identity of the student grade the same test, and where the student's teacher performs a final course evaluation and two external examiners grade a final course exam. The results did not confirm that estimates of gender bias were different in the administrative data, not suggesting that the explanation for the positive gender coefficient found using administrative data is because females perform better at in-class exams or tests measuring different skills. This analysis therefore proceeds to look at other observable determinants of the grade difference using administrative data. Given that estimates of gender bias in the non-administrative bias were similar to the coefficient obtained from the administrative data, there is no reason to mistrust estimates from administrative data based on other student characteristics. In addition, we provided evidence that bias depends on ability. In Table 8, we examined how other observable characteristics are related to grade differences using the administrative data, fixing the coefficient on blind to specific values in the model described in Equation (17). Since the purpose no longer is to compare bias estimates for the same school, year, and course as in the Bergen experiment, we use a sample of recordings from all schools in Bergen for 2008–2015.

[Table 8]

Columns (1)–(3) show the coefficient on the gender dummy with subject- and cohort-interacted fixed effects, and subject-, cohort-, and school-interacted fixed effects. The results are similar to the findings previously shown for 2015. The coefficient changes marginally when moving from Column (1) to Column (2) when adding the school-interacted fixed effect. Middle school attendance is determined by catchment area. This means that the gender balance should be unrelated to school characteristics, since the proportion of females is the same in different types of catchment areas. This may explain why including school-interacted fixed effects only has a small impact on the coefficient. When moving from Column (2) to Column (3), the coefficient on blind score is fixed to −0.12. Holding subject ability fixed significantly increases the size of the coefficient. As we have discussed, this is because the non-blind relationship to subject ability is different than the relationship between blind and subject ability, and females have different ability levels than males.

Columns (4)–(6) repeat the procedure, but jointly include additional observable characteristics. Column (4) suggests that immigrants are positively rewarded by teachers compared to non-immigrants. This is in line with findings in Lindahl (2007) and Falch and Naper (2013). Adding school-interacted fixed effects leaves the estimate unchanged. Column (6) fixes the coefficient on blind score to −0.12, and the coefficient decreases to 0.01 and becomes insignificant. This suggests that, when holding subject ability fixed, immigrants do not receive a positive amount of bias compared to non-immigrants in Bergen.

Column (4) suggests that low-SES students receive a negative amount of bias compared to non-low-SES students. The coefficient becomes larger in magnitude and statistically significant at the 1% significant level when including school-interacted fixed effects. The results suggest that low-SES students are overrepresented in areas where schools are more lenient. Column (6) shows that holding subject ability fixed more than doubles the estimate of the amount of negative bias that low-SES students receive.

Table 8 shows that the estimates of the total amount of group bias, and the total amount of group bias conditioning on ability, parameters described in Equations (15) and (18), may provide widely different estimates of the size of discrimination. According to the econometric model we specify, since we find that $1 - \rho < 0$, estimates of bias that do not take into account subject ability indicates that the bias in favor of the group with lower abilities is larger than when holding subject ability constant.

5 Conclusion

Several studies use data where teachers that know the identity of the students, and teachers that do not, grade students' tests. Systematic differences in grading between these teachers could then be attributed to biased grading. This paper develop an econometric framework that clarifies underlying reasons for differences in grading between teachers that know the students and teachers that do not. In our model, blind scores include subject-specific ability and measurement errors. Furthermore, the model describe that non-blind grades may contain more information than only the subject-specific ability. In addition to subject-specific ability, non-blind includes teacher biased grading according to observable student characteristics of the teacher, the information the teacher has on previous grades and grades in other subjects, and measurement errors. In the administrative data, non-blind may also contain information on subject-specific ability not tested in blind, and the relative performance of students in the non-blind test situations compared to the blind test situation. Our model points to two important issues. First, if administrative non-blind includes more subject-specific ability than blind, or if some students perform better at a specific test type, then using administrative data may not yield an appropriate measure of the total amount of bias one group receives compared to another. Differences across groups can therefore more correctly be ascribed to the effect of test type/grading scheme. Second, if non-blind and blind map differently onto subject-specific skills, the non-blind-blind grade difference is a function of skill. Therefore, differences in grading between the two groups can be a result of different skill levels. This could happen using both administrative and non-administrative data. In addition to developing the econometric framework, this paper compare estimates of bias for comparable administrative and non-administrative data. The results are not able to show that the estimate of the amount of bias females receive compared to males is different when using the two data types. Note that data limitations restrict our conclusion based on this specific dataset.

Furthermore, the analysis provides convincing evidence that the relationship between subject-specific ability and non-blind is not equal to the relationship between subject ability and blind. The consequence of this is important, because it means that subject ability level should be accounted for when estimating the group bias parameter holding the ability level constant.

References

Bertrand, M. & Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, 94 (4), pp.991–1013.

Blank, R. M. 1991. The effects of double-blind versus single-blind reviewing: experimental evidence from the *American Economic Review*. *The American Economic Review*, 81 (5), pp.1041–1067.

Burgess, S. & Greaves, E. 2013. Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31 (3), pp.535–576.

Cornwell, C., Mustard, D. B., & Van Parys, J. 2013. Noncognitive skills and the gender disparities in test scores and teacher assessments: evidence from primary school. *Journal of Human Resources*, 48 (1), pp.236–264.

Deaton, A. 1895. Panel data from time series of cross-sections. *Journal of Econometrics*, 30 (1), pp.109–126.

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

Goldin, C. & Rouse, C. 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. *The American Economic Review*, 90 (4), pp.715–741.

Hanna, R. N. & Linden, L. L. 2000. Discrimination in grading. *American Economic Journal: Economic Policy*, 4 (4), pp.146–168.

Hinnerich, B. T., Höglin, E., & Johannesson, M. 2011. Are boys discriminated against in Swedish high schools? *Economics of Education Review*, 30 (4), pp.682–690.

Hinnerich, B. T., Höglin, E., & Johannesson, M. 2015. Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools. *Education Economics*, 23 (6), pp.660–676.

Landy, F. J., and Farr, J. L. 1980. Performance rating. *Psychological Bulletin*, 87 (1), pp.72–107.

Lavy, V. 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92 (10–11), pp.2083–2105.

Lindahl, E. 2007. Comparing teachers' assessments and national test results: evidence from Sweden. IFAU Institute for Evaluation of Labour Market and Education Policy, Uppsala.

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

Sprietsma, M. 2013. Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics*, 45 (1), pp.523–538.

Van Ewijk, R. 2011. Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30 (5), pp.1045–1058.

Table 1: Institutional details - grading

Region	Dataset	Variable definition	Grader	# graders	Name on test	External/local to school	Test type
Bergen	Administrative	Non-blind	Students' teacher	1	Yes	Local	Course assessment
		Blind	External teachers	2	No	External	National exam
	Non-administrative	Non-blind	Students' teacher	1	Yes	Local	Local test (Tentamen)
		Blind	Another teacher	1	No	Local	Local test (Tentamen)
Rogaland	Administrative	Non-blind	Students' teacher	1	Yes	Local	Course assessment
		Blind	Students' teacher/external teacher	2	No	Local and external	Local exam
	Non-administrative	Non-blind	Students' teacher/external teacher	2	No	Local and external	Local exam
		Blind	External teachers	2	No	External	Local exam

Notes: The table summarize institutional details about the grades used in the analysis. Scores from Bergen are at middle-school level, while scores from Rogaland are at the high school level.

Table 2: Descriptives - Grades

	Bergen				Rogaland			
	Non-adm.	Adm.			Non-adm.	Adm.		
		Same schools	Bergen	08-15		Same schools	Rogaland	08-15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math								
Non-blind average	3.18	3.21	3.66	3.69	3.26	3.00	3.08	3.14
Non-blind sd	1.10	1.20	1.22	1.18	1.27	1.39	1.40	1.37
Blind average	3.08	2.72	3.04	3.25	2.98	2.98	3.07	3.16
Blind sd	1.13	1.24	1.29	1.22	1.32	1.25	1.29	1.27
# Math	39	67	1024	8747	135	649	782	1922
Norwegian								
Non-blind average	3.60	3.95	3.97	3.97	3.16	3.50	3.56	3.52
Non-blind sd	1.01	0.93	1.00	1.00	0.93	1.01	0.99	0.98
Blind average	3.37	3.45	3.56	3.58	2.86	3.21	3.32	3.30
Blind sd	0.97	1.01	1.06	1.02	0.90	0.96	0.96	0.99
# Norwegian	60	38	662	4481	148	536	665	1309
Female	0.43	0.44	0.50	0.49	.	0.40	0.43	0.44
Ses	.	0.22	0.20	0.24	.	0.29	0.29	0.28
Immigrant	0.09	0.08	0.11	0.10	.	0.06	0.05	0.05
# All	99	105	1686	13228	283	1185	1447	3231
Schools	2	2	28	28	15	15	29	29

Notes: Non-blind are grades given by the students' teacher, while blind are grades given by other examinators. Descriptives are on the student level. Columns (1)-(3) consist of students in Bergen exiting middle school in the year 2015 (cohort 1999). Column (4) consists of students in Bergen exiting middle school in the period 2008-2018. Grades are given at the end of the last year of middle school. Columns (5)-(7) consist of students in Rogaland taking high school courses in the years 2010, 2012 and 2013. Column (8) consists of students in Rogaland taking courses in the period 2008-2015. Grades are given at the first and second level of high school. In the non-administrative data, non-blind and blind grades are evaluations of the same test for each student. The test in Bergen is the Tentamen, a locally administered written test. The test in Rogaland is the locally administered end-of-year exam. In the administrative data, the non-blind grade is a teacher evaluation of the students' performance in the course, while the Blind grade is a grade given on a test at the end of the year. Non-blind grades are set before the blind grades are set in the administrative data. In the administrative data from Bergen, the blind evaluation is performed anonomously by two external examinators. In the administrative data from Rogaland, the blind evaluation is set by an examinator together with the students' teachers. This is the locally administered end-of-year exam also used in the experiment. In both Bergen and Rogaland, grades are recorded in the same subject and level.

Table 3: Descriptives - Delta

	Bergen						Rogaland					
	Non-administrative			Administrative			Non-administrative			Administrative		
	W. delta	Norwegian	Math	W. Delta	Norwegian	Math	W. Delta	Norwegian	Math	W. Delta	Norwegian	Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Average delta	0.18	0.23	0.10	0.50	0.50	0.49	0.29	0.30	0.28	0.17	0.29	0.01
SD blind	1.06	0.97	1.13	1.18	1.01	1.24	1.13	0.90	1.32	1.10	0.96	1.25
Average delta/blind SD	0.17	0.24	0.09	0.42	0.50	0.40	0.26	0.33	0.21	0.15	0.30	0.01
Average blind	3.22	3.37	3.08	3.08	3.45	2.72	2.92	2.86	2.98	3.11	3.21	2.98
N	99	60	39	105	38	67	283	148	135	1185	536	649

Notes: The table shows descriptives of the grade difference (Δ_i). Results from Bergen are shown in columns (1)-(6), while results from Rogaland are shown in columns (7)-(12). Columns (1)-(3) and (7)-(9) show descriptives for non-administrative data, while columns (4)-(6) and (10)-(12) show descriptives for administrative data for the same schools, subject, level, and year as in experiment. Recordings in Bergen are from 2015, while recordings for Rogaland are from 2010, 2012, and 2013.

Table 4: Comparing non-administrative and administrative - Bergen

Dependent variable: Grade difference									
	Non-administrative			Administrative					
				Same schools			Bergen		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female		0.009	0.121		0.077	-0.004		0.086**	0.098***
		(0.098)	(0.100)		(0.164)	(0.145)		(0.040)	(0.035)
Intercept	0.182***	0.182***	0.182***	0.495***	0.495***	0.495***	0.537***	0.537***	0.537***
	(0.050)	(0.051)	(0.049)	(0.067)	(0.067)	(0.061)	(0.018)	(0.018)	(0.016)
R2	0.018	0.019	0.113	0.000	0.003	0.202	0.019	0.022	0.254
Adj. R2	0.008	-0.002	0.075	-0.010	-0.017	0.170	0.018	0.021	0.229
N	99	99	99	105	105	105	1686	1686	1686
N Math	39	39	39	67	67	67	1024	1024	1024
N Nor	60	60	60	38	38	38	662	662	662
Female blind	3.609	3.609	3.609	3.286	3.286	3.286	3.571	3.571	3.571
Male blind	2.924	2.924	2.924	2.917	2.917	2.917	3.080	3.080	3.080
Fixed effects									
Subject	x	x		x	x		x	x	
Subject*School			x			x			x

Notes: Results from the non-administrative data from Bergen is reported in columns (1)-(3), while results from the same schools using administrative data are reported in columns (4)-(6). Columns (7)-(9) use administrative recordings from all middle schools in Bergen. Grade differences are recorded in Norwegian and Math. Only recordings from 2015 are included in both the non-administrative and the administrative datasets. Each observation is weighted by the inverse of the proportion of recordings in that subject. The subject weighted average blind grade by gender is shown. The two lowest rows indicate demeaned variables included. The gender variable is also demeaned. * p<0.10, ** p<0.05, *** p<0.01

Table 5: Delta on blind - Bergen

	Administrative				Non-administrative			
Dependent variable: Grade difference (delta)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Blind	-0.25***	-0.22***			-0.11**	-0.09**		
	(0.01)	(0.01)			(0.04)	(0.04)		
School-blind			-0.35***	-0.14***			-0.60**	-0.68**
			(0.03)	(0.02)			(0.27)	(0.28)
Intercept	1.27***	1.17***	1.60***	0.89***	0.52***	0.46***	2.13**	2.38**
	(0.02)	(0.02)	(0.09)	(0.08)	(0.15)	(0.14)	(0.91)	(0.92)
r2	0.16	0.29	0.04	0.20	0.07	0.13	0.07	0.17
Adj. r2	0.16	0.27	0.04	0.19	0.05	0.10	0.05	0.09
N	13228	13228	13228	13228	99	99	99	99
Fixed effects								
Subject*Cohort	x		x		x		x	
Subject*Cohort*School		x				x		
Subject*Cohort*Blind				x				x

Notes: The dependent variable is the grade difference (delta) measured at the individual level. Blind grade is individual blind grade. Only Math and Norwegian recordings are included. School-blind is the school average blind score calculated without the students' own individual blind score for each subject. School-blind to a student drawn in Math is the school average blind grade in Math of all other students drawn in Math at the same school. The regression is weighting each observation with the inverse of the proportion of that subject being recorded. Three lowest rows indicate demeaned variables included. The administrative sample consists of recordings measured in middle school 2008 - 2015, while Non-administrative sample is only for the year 2015. Heteroscedasticity robust standard errors reported. * p<0.10, ** p<0.05, *** p<0.01

Table 6: Delta on blind - Rogaland

	Administrative				Non-administrative			
Dependent variable: Grade difference (delta)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Blind	-0.27***	-0.28***			-0.17***	-0.18***		
	(0.01)	(0.01)			(0.03)	(0.04)		
School Blind			-0.17***	0.02			-0.13**	0.05
			(0.04)	(0.04)			(0.06)	(0.07)
Intercept	0.95***	0.98***	0.62***	0.00	0.79***	0.81***	0.67***	0.13
	(0.04)	(0.05)	(0.14)	(0.14)	(0.09)	(0.11)	(0.18)	(0.21)
r2	0.15	0.35	0.04	0.19	0.11	0.22	0.03	0.22
Adj. r2	0.14	0.25	0.04	0.17	0.09	0.13	0.01	0.13
N	3231	3231	3231	3231	283	283	283	283
Fixed effects								
Subject*Cohort	x		x		x		x	
Subject*Cohort*School		x				x		
Subject*Cohort*Blind				x				x

Notes: The dependent variable is delta (Δ_i) measured at the individual level. Blind grade are individual blind grade. Only Math and Norwegian recordings included. School-blind is school average blind grade calculated without own individual blind grade for each subject. School-blind to a student drawn in Math is school average blind grade in Math of all other students drawn in Math at the same school. Regressions are weighting each observation with the inverse of the proportion of that subject being recorded. The three lowest rows indicate the demeaned variables included. Non-administrative and administrative samples consist of recordings measured in vocational track high schools for 2010, 2012, and 2013 in Rogaland. Heteroscedasticity robust standard errors reported. * p<0.10, ** p<0.05, *** p<0.01

Table 7: Gender bias - fixing ability

Dependent variable: Grade difference									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Girls	0.121	0.121	0.153	-0.003	-0.003	0.042	0.098***	0.098***	0.158***
	(0.100)	(0.100)	(0.097)	(0.135)	(0.135)	(0.128)	(0.035)	(0.035)	(0.033)
Blind		0.000	-0.050		0.000	-0.120		0.000	-0.120
		(.)	(.)		(.)	(.)		(.)	(.)
Intercept	0.182***	0.182***	0.344***	0.495***	0.495***	0.853***	0.537***	0.537***	0.926***
	(0.049)	(0.049)	(0.048)	(0.061)	(0.061)	(0.058)	(0.016)	(0.016)	(0.015)
N	99	99	99	105	105	105	1686	1686	1686
Subject*Cohort*School	x	x	x	x	x	x	x	x	x

Notes: Results from the non-administrative data from Bergen is reported in columns (1)-(3), while results from the same schools using administrative data are reported in Columns (4)-(6). columns (7)-(9) use administrative recordings from all middle schools in Bergen. The columns (2)-(3), (5)-(6), and (8)-(9) show results from a constrained least squareds estimation where the coefficient of blind grade is fixed. Grade differences are recorded in Norwegian and Math. Only recordings from 2015 are included in both the non-administrative and the administrative datasets. Each observation is weighted by the inverse of the proportion of recordings in that subject. Heteroscedasticity robust standard errors reported.. * p<0.10, ** p<0.05, *** p<0.01

Table 8: Group bias

Dependent variable: Grade difference						
	(1)	(2)	(3)	(4)	(5)	(6)
Girls	0.06***	0.07***	0.11***	0.06***	0.08***	0.11***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Immigrant				0.06***	0.06***	0.01
				(0.02)	(0.02)	(0.02)
Low-SES				-0.03*	-0.05***	-0.12***
				(0.02)	(0.02)	(0.02)
Blind	0.00	0.00	-0.132	0.00	0.00	-0.12
	(.)	(.)	(.)	(.)	(.)	(.)
Intercept	0.42***	0.42***	0.86***	0.42***	0.42***	0.86***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	13228	13228	13228	13228	13228	13228
Subject*Cohort	x			x		
Subject*Cohort*School		x	x		x	x

Notes: Results use administrative recordings from all middle schools in Bergen recorded in the period 2008-2015. The table shows results from a constrained least squareds estimation where the coefficient of blind is fixed. Grade differences are recorded in Norwegian and Math. Only recordings from 2015 are included in both the non-administrative and the administrative datasets. Each observation is weighted by the inverse of the proportion of recordings in that subject. Heteroscedasticity robust standard errors reported.. * p<0.10, ** p<0.05, *** p<0.01

Peer Effects from a School Choice Reform

Leroy Andersland[†]

This version: 30 August, 2017

Abstract

In 2005 the city of Bergen, Norway, went from a geographical catchment area high school intake system to a grade point average-based (GPA) intake system. The reform changed the composition of high school peer student characteristics substantially for comparable groups of students before and after the reform. This article compares changes in outcomes for students in Bergen before and after the reform to changes in the outcomes of students in control cities. Positive effects are found on test scores and grades at high school.

Keywords: Analysis of Education, Peer effects, Tracking, High school, Natural experiment
JEL codes: I21, J18

[†] Department of Economics, University of Bergen, 5020 Bergen, Norway, email; leroy.andersland@econ.uib.no

1 Introduction

A significant and ongoing debate in educational research and policy concerns how to place students into groups to enhance the learning environment. One important strand of the literature has studied the optimal mixture of a student group in field experiments (Booij et al. 2017; Carrell et al. 2013; Duflo et al. 2011). Do high-ability students benefit from having high-ability peers, or is it preferable to have a mixture of high- and-low ability peers? The answer to this question may vary by student age, whether cognitive or non-cognitive outcomes are studied, type of peer ability, subject, and institutional setting.

A main argument for having a system with similar students grouped together in schools is to allow the teacher to tailor content and pedagogical techniques to a homogeneous group. Another argument is that it can potentially improve student interaction. For example, high-ability students may create a culture for learning or push each other toward better achievement when grouped together. The main arguments for mixing students with different backgrounds rely on the notion that students may gain from being part of groups that include students that are diverse in various ways. High-ability students may gain from ranking in the top of their class, while low-ability students may gain from interacting with high-ability students. There are, however, many ways of grouping students, and it is possible to do so using different dimensions. In the end, the question of which system and what kind of mechanisms is most preferred are empirical.

Evidence regarding peer effects on students in their natural environments relies on finding natural experiments; the most frequent types used in the literature include housing vouchers, busing students to different schools, natural disasters, and school acceptance cutoffs. This paper adds evidence to the literature by using a school choice reform that varied peer characteristics at the school level. Specifically, we examine a reform in Bergen, Norway's second-largest city, which changed its high school intake from a catchment area approach to a performance-based intake system. High schools located in the city center (downtown) were more attractive to students than high schools outside the city center (suburbs). When the intake system was neighborhood-based, students living downtown attended downtown high schools. Once the intake system became performance-based, most high-ability students living downtown still attended central schools, while many lower-ability downtown students were shut out of downtown schools by high-ability suburban students. In sum, the reform decreased variation within high schools, and increased variation across high schools in Bergen.

This paper compares the outcomes of students in Bergen before and after the reform to comparable students in other cities to uncover their response to changes in peer characteristics. High-ability downtown students attended high schools where the average peer student middle school GPA increased by 0.65 over the middle school GPA standard deviation (SD) from before the reform to after it, compared to similar groups in other cities. This is equivalent to going from a school at the median to a school among the top 10% in both Bergen and control cities before reform. The results show positive and significant effects on centralized, externally evaluated exams in some subjects for high-ability students as a consequence of the reform. For lower-ability students, the reform implied attending high school with less variation in peer characteristics. Consistent with recent findings from field experiments (Boiji et al. 2017, Carrell et al. 2013, Duflo 2011), our results suggest that high school performance for these students increased as a consequence of the reform. The intake reform led to a natural experiment that generated a type of tracking similar to that achieved in experiments. The reform makes it possible to identify effects on high ability students of changing peers from mixed to high ability (high-high). For low ability students it is possible to find effects of changing peers from mixed to low (low-low). The effects of this type of tracking are relevant in cases where one decides between dividing a group based on prior ability or not.

Furthermore, the analysis suggests that students obtaining the highest grades before and two years after the reform were not necessarily comparable for two separate reasons. First, the average middle school GPA increased for cohorts completing middle school after the reform, largely because the pre-reform middle school GPA was of relatively low importance, since it had very few actual consequences for the students. After intake reform, middle school GPA became the main measure that determined which high school students would attend, so many students showed increased efforts in middle school to achieve the higher grades that would allow them to attend the school of their choice. Second, the predetermined background characteristics of the highest-scoring students at middle schools were different before and two years after the reform, which can only be interpreted as meaning that highest-scoring students are different under low- and high-stakes systems. This insight adds another layer of complexity in finding a comparable group of students in Bergen before and after reform. This analysis therefore employ a difference-in-difference-in-difference (DDD) strategy that uses the fact that high-ability middle school students in suburban areas in Bergen were exposed to the same reform, but experienced a much smaller average change in peer characteristics in high school than high-ability downtown students. The findings from this empirical strategy are weaker, but still support the hypothesis of the existence of some positive effects of the reform for this group.

The main contribution of this paper is to introduce a new type of natural experiment to identify the effect on educational outcomes of significantly changing the peer environment. While changing the characteristics of the peers, we attempt to keep school type and travel distances fixed for a particular group. In addition, this new identification strategy permits an investigation of the consequences of tracking for different groups, and opens novel perspectives on mechanisms that are explored in detail using high-quality register data. Section 2 reviews the literature and places this paper's contributions in the context of existing scholarship, while Section 3 describes the institutional realities of high school intake in Norway. Section 4 explains how the empirical design was implemented to exploit the exogenous variation in peer characteristics created by the reform. The baseline sample was constructed from administrative records. Section 5 details the data that are important for interpreting the results. Section 6 contains the results, including a discussion of mechanisms and robustness checks, while Section 7 summarizes and concludes the paper.

2 Literature review

Manski (1993) formulated three concepts that are fundamental for the understanding of peer effects. The first is the endogenous effects—students tend to change behavior according to the behavior of the group. The second is exogenous effects, which expresses that students current behavior depends on the groups background characteristics. Finally, the correlated effects addresses the fact that student tend to select into groups based on unobserved characteristics. The reflection issue arises because of the existence of a possible multiplier in that peer behavior can affect one's own behavior, which in turn can affect the peers. Most modern papers address the identification problem stemming from endogenous peer group formation and reflection, but only a few are able to separate exogenous from endogenous effects.

Hoxby (2000) authored one of the first studies that explicitly addressed the selection issue across schools by using exogenous variation in peer characteristics within schools and across years. The results indicated positive peer effects of high-ability peers, which were stronger in a same-race context for primary schools in Texas. Hanushek et al. (2003), Betts and Zau (2004), and Lavy et al. (2012) reported similar findings. Using Norwegian data, Black et al. (2013) studied long-run outcomes such as IQ scores, teenage childbearing, educational choices, and adult labor market status and earnings. The study found positive effects among girls of having more females in the cohort. No effect was found due to variation in average education of the mother, but average income of the father appeared to play a role in students' long-term outcomes. Other results from Norway using the same method are found in Bonesronning (2008) and Boenesronning and Haraldsvik (2014). These studies showed that school achievement was negatively affected by the presence of classmates from dissolved families and students with less-educated parents respectively.

Carrel et al. (2013) used variations in squadrons' standardized test scores as the peer variable to identify peer effects at the Unites States Air Force Academy. This analysis suggested a positive effect of peer Scholastic Aptitude Test scores on freshman GPAs among low-ability students. Building on these results, a follow-up study was conducted in which low-ability students were randomly assigned to squadrons with high-ability peers. The resulting significant negative effects for low-ability students and lack of effects on high-ability students were taken as at least partial evidence of the importance of endogenous, within-squadron peer group formation. Angrist (2014) cited this study as evidence that the standard approach of regressing outcomes on peer means, with variation mainly coming from naturally occurring variation, is not reliable.

The alternative to using naturally occurring variation is to conduct randomized experiments that manipulate the peer characteristics of individual students. As in Carrell et al. (2013), Duflo et al. (2011) manipulated peer groups in an experimental setting by streaming students into ability groups. Low-ability students were put in groups with other low-ability students, while high-ability students were put into groups with other high-ability students. The results indicated that all students benefited from tracking, including the low-ability students assigned to low-ability groups. The researchers concluded that these results show that students benefit when teachers are able to adjust their teaching approaches to a homogenous class. Booij et al. (2017) randomized students into tracked groups using an expanded set of track combinations, so they were able to look at the effects of different combinations of group compositions and find results consistent with both Carrell et al. (2013) and Duflo et al. (2011).

2.1 Natural experiments

Many articles on peer effects in schools employs some kind of natural experiment to identify those effects. A frequent choice is policy interventions that are intended to desegregate neighborhoods or schools. Examples of this from the United States are Moving to Opportunity housing vouchers (MTO) and Metropolitan Council for Educational Opportunity (Metco). Kling et al. (2007) based their study on the fact that a lottery assignment mechanism was used to assign MTO vouchers to families; these vouchers gave families the opportunity to move to lower-poverty areas. Comparing families that were offered vouchers to those that were not, the researchers did not find any effect on adult economic self-sufficiency. They did however find beneficial mental health effects for female youth that were offset by negative health effects for male youth. Kling et al. (2005) used the same policy intervention to study criminal behavior and found similar results; female criminality went down when moving to lower-poverty areas, while the effects for males were more mixed. Ludwig et al. (2013) studied long-term outcomes and found results consistent with previous research. Chetty et al. (2016) employed newly available data on children younger than 13 at the time of random assignment. Restricting their sample to this cohort, they found significant positive effects on earnings for all groups in their mid-twenties.

Angrist and Lang (2004) analyzed the effect on test scores for students in suburban schools that received a fraction of new “Metco students” from low-income areas. Metco is a desegregation program that sends low-income students out of poor Boston districts into schools in the surrounding suburban areas. The study did not find substantial effects on students already attending those schools. Sacerdote (2011) notes that the literature generally shows modestly positive effect on academic achievement gains, but that the effects on non-academic outcomes appear to be much larger. Sacerdote (2001), Duncan et al. (2005), DeSimone (2007), Wilson (2007), Kling et al. (2005), Kling et al. (2007), and Carrel et al. (2008) looked at outcomes such as drinking, smoking, cheating, sexual activity, criminal involvement, health, and racial attitudes. Recently, Rao (2015) studied variations in the proportion of poor children in Indian middle class schools and found that overall attitudes towards the poor became more altruistic.

Another type of natural experiment uses the regression discontinuity framework, studying students that apply to selective high schools; some are accepted and some rejected based on admission scores. Employing this strategy, Clarke (2010) found only small effects on test scores of attending selective UK schools. Jackson (2013) used this design with single-sex schools in Trinidad and Tobago, while Abdulkadiroglu et al. (2014) studied the public school systems in Boston and New York. Both show that students just above the admission cutoffs attend high schools with students that score about on average 0.5 standard deviation higher on a predetermined test than students right below the cutoff. Jackson (2013) only found effects for a group of students that had expressed strong preferences for attending selective single-sex schools and some negative effect on selecting science courses, while Abdulkadiroglu et al. (2014) found no effect of attending elite schools on students near the cutoff for admission to these schools.

One recent study used an empirical design to avoid the issues of endogenous peer groups, correlated effects, and reflection. Dahl et al. (2013) examined social interactions in program participation. Their results showed, among other things, the importance of naturally occurring preexisting peer groups, as they were able to identify effects on siblings and coworkers on program adoption by varying the "price" of the social program. Translating this to a school setting means that even though a study is able to control for endogenous sorting across schools or classes, the endogenous sorting within schools or classes may be an important point of focus for studies analyzing school-situated peer effects. Other analyses using the partial population approach are found in studies of the PROGRESA program in Mexico. The PROGRESA program provided cash incentives for parents to send their children to school. Peer effects can then be identified on ineligible children that are in the naturally occurring peer groups of eligible children. Angelucci et al (2009), Bobonis and Finnan (2009), and Lalive and Cattaneo (2009) all found substantial positive peer effects on school attendance. To the best of our knowledge, however, no studies have used this method to identify peer effects on academic achievement.

Empirical designs using natural experiments are often unable to separate neighborhood or school effects from peer effects. There may be other differences between high- and low-poverty neighborhoods than resident incomes, and there are other differences between elite schools and other schools than their students. Designs of the type found in Angrist and Lang (2004) explicitly address this by focusing on the effects on students who were already attending schools that experienced a change in student composition. Besides employing a new type of natural experiment, a key contribution of the present study is that it keeps variables such as school type and travel distance fixed, while varying average peer characteristics substantially. The combination of these two features is not often found in the scholarly literature. In addition, the reform that changed the high school intake from a geographical catchment area based system to a GPA based intake system is similar to switching from ability mixing to tracking. Therefore, we contribute with a natural experiment that allows us to explore the effects of ability tracking at high school for both low and high achieving students.

3 Institutional setting

Children in Norway start school in August of the year in which they reach six years of age. Children normally attend primary school until age 14 and middle school from age 14 to age 16. Most primary and middle schools are public, with intake based on geographical catchment areas. When students finish middle school, most students choose between applying for academic or vocational tracks at high school. Around half of students choose to start on the academic track, with about 75% of that group graduating within three years.

3.1 High school intake

In 2005, 95% of high school students in Norway attended public high schools. High school intake systems are regulated at the county level; private high schools have separate mechanisms to accept students. The approaches adopted can be divided into middle school GPA-based intake systems, geographical catchment area-based intake systems, and combinations of the two models (Brugård 2013). Bergen, where the reform examined in the present study occurred, is located in Hordaland County. Before the school year starting in 2005, Bergen had a system by which most students completing middle school were assigned a high school by the county school administrative office, which was guided by rules that obliged them to divide all middle school students into GPA groups, and then divide these students among high schools so that each high school had a roughly equal proportion of students from each GPA group.[1] One reason for this approach was to try to avoid the development of "good" and "bad" schools. In practice, students were generally assigned to

[1] Source: Nils Skarvhellen, Head of Intake office at Hordaland Fylkeskommune. (20.10.2015), Knudsen, Sortevik and Woldset, Government proposal analysis (2003)

schools was close to their homes to reduce travel time; there was some flexibility on the administrative office's part to deal with students with strong desires to attend a certain high school. The Hordaland school administrative office noted that this system required significant effort on their part. The demands of the intake system, combined with increasing pressure from different interest groups, were the reasons that Hordaland County changed its intake system in the school year beginning in August 2005 to a middle school GPA-based intake system. The county government passed the rule change in October 2004. In the next high school intake students could list schools based on preference, and were accepted to their first choice if their middle school GPA was above that school's admission level. Each school had a limited number of seats, so the admissions level varied depending on the number and middle school GPA of the applicants. Only the central school authority at the county level is involved in the acceptance procedure and not the schools.

The control municipalities used for empirical comparison purposes are Trondheim, Stavanger, Drammen, and Kristiansand, four of the five other largest cities in Norway. Oslo, the capital and largest city, is not included among the control cities because a separate school choice reform took place at the same time as in Bergen.[2] Drammen had a catchment area-based intake system, while Trondheim, Stavanger, and Kristiansand all based intake on middle school GPA.

3.2 Curriculum and grading

Learning structure and course compositions at schools are regulated at the national level in Norway. This means that all students who attend a public school have access to approximately the same range of courses and attend schools with similar learning principles

[2] We chose to focus on the reform that happened in Bergen since the reform there was a total transition, while the reform in Oslo was only a partial change in intake systems.

and goals. For the school year 2006–2007 a reform in learning principles took place, called "Kunnskapsløftet". The most important changes at the high school level were changes in course composition; students were not differentially exposed to the reform within cohorts, only across cohorts.

4 Empirical design

A school intake reform may significantly change the composition of students at high schools. With a catchment area intake system, students generally attend the geographically closest high school. With a performance-based intake system, however, high schools consist of students who apply and are accepted to each high school based on middle school performance. The degree of change in the composition of students after an intake reform will depend on the attractiveness of the high school. A change from a catchment area to a performance-based system will lead to a negative selection of students at less attractive schools, since high-ability students will have the option to leave, which most low-ability students will not. The same change will also lead to a positive selection into attractive schools, since high-ability students from outside the catchment area will be chosen over low-ability students from within the catchment area.

Comparable students before and after an intake reform may end up attending a high school with very different peers. In this paper we use these changes to analyze peer effects. The first step is to find comparable students before and after the reform for which the reform's main effect was changing the characteristics of their high school peers. Besides peer characteristics, a school intake reform can change both daily travel distance and the type of high school for comparable students. The group for which reform is most likely to mainly change peer characteristics are high-ability students living in the catchment areas of attractive schools; they would have attended those schools before reform, and because they still qualify and there is no obvious reason for them to apply to less attractive schools, they are likely to continue attending those same schools. We first chose to focus on high-ability students living in attractive schools' catchment areas, since we expect that this group should experience a significant change in peer characteristics without any changes in distance traveled and type of school.

High-ability students belonging in downtown middle school districts in Bergen before the reform attended high schools located in downtown Bergen; most students average high school peer ability as measured by middle school GPA was close to the average of Bergen as a whole. After the transition to a GPA-based system, they still attended those same downtown high schools, but now many of their low-ability middle school peers were replaced by high-ability students from the suburbs. To identify the effect of these changed peer characteristics, we compared the change in outcomes of high-ability downtown students in Bergen to the change in outcomes of high-ability downtown students in other large cities in Norway in a difference-in-difference (DD) setup. Model 1 is defined as:

$$Y_{it} = \beta_0 + \beta_1 BERGEN_{it} + \beta_2 REFORM_{it} + \beta_3 (BERGEN_{it} * REFORM_{it}) + \beta_4 X_{it} + \varepsilon_{it} \quad (1)$$

Where i denotes individual student, t denotes cohort, Y_{it} are outcomes that can be affected by changed peer characteristics, $BERGEN_{it}$ is a dummy variable (1 for high-ability student living downtown, 0 if the student has high ability and lives in another city center). High-ability students are defined as those having middle school GPAs in the top 25% of their citywide cohort, while downtown students are those who attended middle school in the downtown area of the city.[3] Middle school attendance is almost exclusively determined by a middle school level geographical catchment area. $REFORM_{it}$ is a dummy indicating 1 for the cohorts

[3] We vary the GPA threshold in the robustness section.

applying to high school after Bergen's reform (born from 1989–1991), and 0 for cohorts applying for high school before it (born 1986–1988). X_{it} is a vector of individual-level control variables and middle school dummies. Individual level controls are parents' earnings, parents' years of education, and gender.

[FIGURE 1]

Figure 1 shows the development of average incoming peer GPA at high schools for each cohort for high-ability students in downtown Bergen and control cities. Average peer GPA remains stable at around 4.4–4.6 for both groups before reform. Peer GPA is slightly higher in the control groups, something that could be explained by the fact that three of the four control cities had a GPA-based intake system in the period studied. The reform was implemented for the cohort born in 1989, and we saw a sharp increase in peer GPA in the treatment group for this cohort, which stabilized at a higher level for the subsequent two cohorts. There was no change for the control cohorts. Peer students' middle school GPA increased by 0.65 of one SD of middle school GPA from before to after reform, compared to comparable groups in other cities. This is equivalent to going from a school at the median to a school among the top 10% in both Bergen and the control cities before reform.

5 Data and variable definition

Data were taken from Norwegian administrative records. Middle school grade information is available through a centralized middle school database, while information on middle school and high school attendance is available through education records detailing the schools and tracks that individuals attend and complete. High school grade information is available from two sources, the school administrative grade records (a database with grade information collected from the various school administrative systems) and the Norwegian certificate database, which contains high school grade information for all certificates granted. Both databases are used in this analysis because each has strengths and weaknesses. The certificate database contains only grade information on those who complete their certificates at high school, so dropouts' grades are not present. The certificates database however has grade information for more cohorts than the administrative database.

The baseline sample is students who started high school immediately after middle school.[4] Data on school absence is available from the certificates database for those who completed school. This measure comes from teachers' recording the number of hours and days a student was absent from class during the school year.

The data allow us to link students and parents, so we can use parents' years of education and yearly earnings as control variables; these are both measured when the students are 10 years old. Earnings are measured in 1996 NOK.

[TABLE 1]

Table 1 offers descriptive statistics of the sample; the baseline sample consists of high-ability downtown students in Bergen and similar students in Kristiansand, Stavanger, Trondheim, and Drammen. The table is divided into three panels: panel a) show descriptive statistics of covariates, panel b) shows the peer GPA variable, and panel c) shows descriptive statistics of high school outcomes. Column (1) shows the pre-reform means of the treated group, Column (2) shows the SD of that group. Columns (3) and (4) show the difference in means between treatment and control before and after the reform. Column (5) shows the number of observations of each variable. The top 25% downtown students in Bergen, Kristiansand,

[4] We have verified that the reform did not affect applications and intake to academic tracks.

Stavanger, Trondheim, and Drammen give a total baseline sample of 1869 students. Missing observations on covariates are dealt with by replacing them with the value 0 and including a dummy for the missing observation of the covariate.

From the descriptive statistics of covariates we note that students belonging to the treatment group have a lower proportion of females and parents with higher earnings and more education. These differences are somewhat smaller after reform. Peergpa is constructed from the average incoming middle school GPA of the peers of the high school students. We note that before the reform, high-ability downtown students in Bergen on average attended schools with lower-ability peers than in other cities. This is due to the intake system in Bergen pre-reform not being performance-based, while three of the four cities in the control cities did have performance-based intake systems. Comparing the difference between columns (3) and (4) in panel b) shows that the treatment group increased their average peer GPA at high school by 0.41 compared to the control group.

Firstyear GPA is the average grade for the first year of high school, while "High School GPA" is the average of all grades in high school. These two measures largely contain grades assigned locally by the teacher. Absence days and Absence hours are the number of days and hours of recorded absence during all years at high school. Absence days are the number of full days that a student was not recorded as present in any class at school. Absence hours are the number of hours of recorded absences from class, not including full-day absences. "Select basic math year 1" indicates whether the student selected the less advanced math course in the first year of high school.

Norwegian exam in year 3 is a compulsory national exam at the end of high school that is externally administered and graded. Norwegian II exam year 3 is the second formal written language for the student. Students decide themselves which written language is their main language. For more than 90 % of the students, the Norwegian exam is “Bokmål”, while the Norwegian II exam is “Nynorsk”.

Exams in years 1, 2, and 3 make use of the fact that students are randomly drawn to take exams in different subjects. Scores in different subjects are pooled for each student by year since the number of students drawn for each subject is relatively small. Only about 30% of students are drawn to take an exam in year 1 in any subject, which explains the smaller sample size. The advantage of the exams in year 1 is that the exam-takers are randomly drawn among mandatory subjects, which means that that the coefficient are not inadvertently capturing mechanisms that involve a change in course composition. A larger proportion of the students take a standardized written exam in years 2 and 3, which are drawn among electives.

6 Results

The first results using Model 1 are shown in columns (1)–(3) in Table 2. The table focuses on the high-ability downtown students who experienced a large increase in peer ability. Column (1) presents the results without any controls, while Column (2) add middle school dummies. Column (3) shows the preferred specifications were controls for background characteristics and middle school dummies are included. The focus of the discussion will be on the empirical specification including middle school dummies and background characteristics. Each row gives the estimate of β_3 from Model 1 with the dependent variable indicated in the row header. For now, β_3 is interpreted as the average treatment effect on the treated (ATT) of the intake reform on high-ability downtown students. Section 7 goes into detail about what may explain the findings.

[TABLE 2]

A positive coefficient is found on both Firstyear GPA and High school GPA, but the effect is only significant on average grades in the first year of high school after including middle school dummies. After including both middle school dummies and background characteristics, the coefficient on Firstyear GPA is 0.10. There is a negative though not significant coefficient for total hours absent from high school and a positive insignificant coefficient for total days absent. Not finding any effect on absence is in consistent with travel time being unaffected by the reform for these students. The next row shows a non-significant increase in the likelihood of selecting a basic as opposed to an advanced math course during the first year of high school. Thus, the results does not give any conclusive evidence of whether high-ability peers encourage more advance course taking, or if it makes it more difficult to get a seat at a limited number of spots at these courses.

The effect on Norwegian exam in year 3 shown in Column (3) are statistically significant at the 10% significance level. The size of the effect is 0.20 and stable across different specifications. The results show that Nynorsk also increased by 0.30. This effect is significant at the 5% level. Even though there are fewer observations of Exam year 1, a statistically significant positive effect at the 1% significance level is found on this measure. The size of the coefficient is 0.48, and is the largest effect on the achievement measures shown in Table 2. In sum, together with the effect on GPA, the effect on national exams suggests the reform's positive effect for the high-ability downtown students. The effects on Exam year 2 and 3 are not significant in any of the specification. A possible explanation is that these exams are drawn among elective courses, which may be affected by the reform, making this exam measure more sensitive to mechanisms that cause students to change course composition at high school.

[FIGURE 2]

Figure 2 plots average outcomes by cohort, allowing for a graphical inspection of how outcomes change over time. Solid circles and triangles indicate averages, while 95% confidence intervals of means are indicated with crosses. Confidence intervals are tighter for the control group since that group is larger. Norwegian exam scores and GPAs averages move relatively coherent before reform, with a trend shift for the treatment group at the time of the reform. This supports the assumption that if the reform had not happened, the two groups would have had the same development in outcomes.

6.1 Reform effects on all groups in Bergen

Table 3 show subsample estimates of β_3 in Model 1 on all groups of students in Bergen. All estimations are performed with middle school dummies and background control variables. Columns (1)–(4) show estimates for the low-, medium-low-, medium-high-, and high-ability downtown students, while columns (5)–(8) show these results for suburban students across the same achievement groups.[5] The first row shows that higher-ability students received higher-ability peers after reform, and that this effect was strongest for downtown students. For low- and medium low-ability students peergpa increased much weaker or not significantly at all. Given the large increase in peergpa for high-ability students, a larger fall in peergpa for other groups of students could be expected. The reason for this is that average middle school GPA in Bergen increased after the reform. Further discussion of the consequences of this fact

[5] Students are split into equal sized groups within their cohort and city based on where they ranked on the middle school GPA distribution.

appears in section 7.1. Another consequence that can be inferred from the first row is that the variation in student characteristics within schools decreases.

[TABLE 3]

The second and third rows show clear positive and significant effects on average high school grades for all groups of students, except for high-ability students. These results show that the intake reform benefited these other groups of students. There is a negative coefficient on hours absent significant at the 5% level for downtown students with medium-low ability, while there is a positive significant coefficient at the 10 % level on days absent for low-ability suburban students. Thus, the effects on absence are inconsistent.

The effect on the centralized exams in the third year is most pronounced for high-ability downtown students. The positive effects on GPA that appeared for the other groups of students are not found to the same degree on exam scores. One explanation for this is that, as measures of academic achievement, the exams are subject to more noise. Alternatively, the school grading captures improvements in abilities that are not measured in exams. An example of this is classroom behavior.

Regarding the effect on the pooled exam score measures, a larger positive effect on first-year exams for high-ability downtown students are found than for most of the other groups. No significant effects are found on second-year exams, while on third-year exams a significant negative effect for high-ability suburban students and a significant positive effect for low-ability suburban students. As noted above, the first-year exam involves fewer students, while second- and third year exams are vulnerable to potential mechanisms that cause students to change course selection because of the reform. This may explain why the findings for these measures are less coherent. Average outcomes by group over cohort can be visually inspected in Figures A.2 and A.3 in the Appendix.

6.2 Placebo and Robustness

β_3 identifies the ATT effect of the reform on downtown students ranked in the top 25% of their cohort in Bergen, with the assumption that without the reform, they would have had the same trend in outcomes as downtown students in control cities ranked in the top 25% of their cohort and city. To determine if that would have been the case, trends before the reform are examined. This is possible since we observe outcomes for three cohorts of students before reform.

[TABLE 4]

Figure 2 allows for visual inspection, but for a formal test Model 1 is estimated with the adjustment of keeping only the three cohorts before reform and defining two placebo reforms starting in school years 2003–2004 and 2004–2005. Results are shown in columns (1) and (2) of Table 4. Only two of the 22 tests gave a significant coefficient at the 10% significance level. This does not provide strong evidence against the common trend assumption, though one weakness of this test is that the lower sample size offers less precision.

[TABLE 5]

Table 5 show how effects change when the estimation sample or model specifications change. Column (1) gives the baseline estimates reported in the main results. Column (2) shows estimates in a sample containing downtown students scoring in the top 33% of their city cohort. Coefficients on Norwegian and Norwegian II exams remained positive, but they were no longer significant. One possible explanation is that it was students with the highest ability that gained the most from a change in peers. Column (3) includes high-ability students from more cities than those included in the baseline sample. Both the effects on grades and test scores became less noticeable, though this could be because this sample definition are comparing trends in groups that were less equal than the groups compared earlier. Columns (4)–(6) explore how the estimates are sensitive to adjusting the cohorts included in the sample. Standard errors on effects increased when reducing the sample size to include only cohorts closer to reform, while the effect on exams stay significant.

To correct for possible intragroup correlation in error terms, standard errors were clustered at the high school*year level. Table A.1 in the Appendix shows that the results are somewhat sensitive to the level of clustering. Standard errors on test scores decreased when clustering on middle school or city. One possible reason is few clusters; there were 25 middle schools, while there were five cities.

[FIGURE 3]

The last robustness check performed is based on the permutation method proposed in Buchmueller, DiNardo, and Valleta (2011). We have assumed that the policy change happened in each of the 20 largest cities in Norway (excluding Bergen and Oslo), and estimated DD coefficients for the top 25% students in each of these cities and for the other three ability groups in each city. Figure 3 shows the distribution of these coefficients and the 95th percentile in the distribution. By comparing the coefficient for Bergen to the empirical distribution of DD coefficients for the other groups, we rejected or kept the null hypothesis of no effect in Bergen for each outcome. The results from this robustness check show that none of the coefficients are above the 95th percentile in the distribution of coefficients. A way to increase precision is to pool the two Norwegian exam scores and inspect inference is for this measure. Pooling Norwegian scores results in nearly a doubling of the sample size, and results are significant at the 1% significance level with robust standard errors, and above the 95ht percentile in the distribution created with the permutation test. Bergen is the second largest city in Norway. Choosing fewer and larger cities decrease the variation in the distribution of coefficients. In total, results from the permutation test do not suggest intragroup correlation in error terms lead to too small standard errors.

7 Mechanism

7.1 Testing for selection and "the incentivizing effect"

The 25% best students from downtown Bergen districts and the 25% best students from downtown districts in the control cities are student-group categories that students could switch in or out of because of reform. The ATT of a school choice reform on high-ability downtown students would be biased if the top 25% downtown students were different under a catchment area system and a GPA-based system, as for example if students changed their catchment area to one with their preferred school as a response to the introduction of a catchment area system. This could prevent the treated group from being comparable before and after reform. Machin and Salvanes (2010) showed that house prices in Oslo remained sensitive to school intake reform that took place in 1997, even a decade later. This may be less of an issue in Bergen since the system before the reform was not strictly based on catchment area.

A similar situation would arise if students ranked in the top 25% of their cohort were not the same before and after reform. Haraldsvik (2014) studied the effect on middle school grades of students in Bergen as a consequence of reform. The study revealed that those grades increased in the district as a whole with the transition from the catchment system to the GPA system. Haraldsvik proposed that a performance-based system incentivized some or all students to work for better grades in order to increase their chances of attending their preferred high school.

The reform was announced in the fall the year before it was implemented. For the first cohort applying to high school after the reform, the adjustment time was less than a school year. The first cohort after reform should also have been less incentivized to increase their grades, since there were fewer observable differences between high schools. The second and third cohorts had more time to adjust to intake reform, and the differences between schools would have been more evident.

[TABLE 6]

One way to test whether there was selection of students into the treated group due to reform is examining changes in predetermined background variables, and if students were incentivized by the reform, it would be revealed by grades determined before high school. To test for selection and the incentivizing effect, Model 2, a modified version of Model, 1 is implemented:

$$Y_{it} = \alpha_0 + \alpha_1 BERGEN_{it} + \alpha_2 Cohort1989_{it} + \alpha_3 Cohort1990_{it} + \alpha_4 Cohort1991_{it} + \alpha_5(Cohort1989_{it} * BERGEN_{it}) + \alpha_6(Cohort1990_{it} * BERGEN_{it}) + \alpha_7(Cohort1991_{it} * BERGEN_{it}) + \alpha_8 X_{it} + \epsilon_{it} \qquad (2)$$

The main change from Model 1 is that the Bergen indicator now is interacted with indicators for each post-reform cohort. The dummies $Cohort1989_{it}$, $Cohort1990_{it}$, and $Cohort1991_{it}$ are the indicators for the three post-reform cohorts.

Table 6 show estimates of α_5, α_6, and α_7 in Model 2. Panel a) shows results when the dependent variables are background characteristics and middle school GPA. Panel b) show results when dependent variable are high school peergpa and measures of academic performance in high school. There are generally smaller differential changes in predetermined background characteristics for the first cohort than for the rest. The second cohort shows larger differences, while for the third cohort there were negative significant coefficients on fathers' and mothers' years of education. Coefficients on parents' income were negative but insignificant, while a positive insignificant coefficient appeared for the female dummy. These results confirm a hypothesis of dynamic response to school choice reform. Students scoring in the top 25% of their cohort in the city center of Bergen were different before the reform and two years after the reform.

Table 6 shows that middle school GPA increased. This finding could be explained by the "incentivizing effect" that the school choice reform had on student middle school grades. The top 25% of downtown Bergen students had higher middle school grades after the reform, a finding that is in line with Haraldsvik (2014). The second main explanation for the ATT of the reform is therefore that high-ability students in Bergen became better because of the reform before entering high school.

Panel b) focuses on high school academic outcomes with the new specification. It indicates that selection into the top 25% of downtown students in Bergen affected the high school outcomes of this group. The results in Table 2 show that including background variables did not change the results significantly. However, if the significant coefficients in panel a) indicated changes in unobservable factors, that would suggest that the DD coefficient is a lower bound on the effect of the reform. The incentivizing effect indicated by the positive effect on middle school GPA is not an issue if it was a transitory shock to abilities that do not affect outcomes in high school. If it was not a transitory shock in abilities, then part of the observed effect of the reform could be explained by this phenomenon. The next section specifies a model that is designed to take into account both selection into the group of the top 25% of students in Bergen and the incentivizing effect due to school choice reform.

7.2 Accounting for selection and direct effects

A school choice reform could affect the high school outcomes of the top 25% of downtown Bergen students through channels other than a change in peer characteristics in high school. The ATT would then not only reflect a peer effect but also these alternative mechanisms. The first of the two main alternative explanations indicated in the last section is that the top 25% of students in Bergen were not the same before and after reform. The second is that the top 25% of downtown Bergen students had higher ability after the reform because they studied harder at middle school in order to be accepted into selective high schools.

[TABLE 7]

One way to separate the peer effect from these explanations is employing the fact that all students in Bergen underwent the school choice reform, but not all of them experienced the same change in high school peer characteristics. As shown in Column (8) in Table 3, the top 25% of suburban students did not experience the same change in peer characteristics, even though they were equally subject to the reform. The procedure would then compare changes in test scores between downtown students in Bergen and in control cities to changes in test scores between suburban students in Bergen and in control cities. Changes in group composition and the incentivizing effect would no longer pose a concern if these effects were equal in the downtown and suburban areas. To implement this strategy, the following specification (Model 3) is estimated:

$$Y_{it} = \gamma_0 + \gamma_1 BERGEN_{it} + \gamma_2 REFORM_{it} + \gamma_3 DOWNTOWN_{it} + \gamma_4(BERGEN_{it} * REFORM_{it}) + \gamma_5(BERGEN_{it} * DOWNTOWN_{it}) + \gamma_6(REFORM_{it} * DOWNTOWN_{it}) + \gamma_7(BERGEN_{it} * REFORM_{it} * DOWNTOWN_{it}) + \gamma_8 \mathrm{X}_{it} + \varepsilon_{it} \qquad (3)$$

Columns (1) in Table 7 show estimates of γ_7 in Model 3. First we note that since both downtown and suburban students experienced an increase in average peergpa, the relative effect on peergpa goes down. The coefficient of GPA measures are about the same, but significance disappear for Firstyear GPA. The coefficient for Norwegian exam year 3 is 0.16 and insignificant, while the coefficient with controls on Norwegian II is 0.34 and significant at the 5 % significance level. Significance disappears for Exam year 1, while the coefficient change sign for Exam year 3. The disappearance of significance for some of the outcomes could be explained as a direct effect of the incentivizing effect of the reform. Alternatively, the disappearance could be explained by the lower relative increase in peergpa using this strategy. The coefficient for Norwegian II suggests that high-ability downtown students still gain from the reform relative to high ability suburban students.

[FIGURE 4]

To test for whether the DDD strategy accounts for the selection and the incentivizing effect, Model 3 is estimated with background characteristics and middle school GPA as dependent variables. Table A.3 in the Appendix shows that the significant effect from middle school

GPA disappears, while there is still a negative effect on maternal years of education. This means that we cannot reject a null hypothesis that there was no incentivizing effect, while there may still be some unaccounted-for selection in the model using the DDD identification strategy.

7.3 School effects

The reform in Bergen allowed students to choose which school to attend. For at least 10 years before reform, a catchment area design in which students' ability as measured by middle school GPA was used to distribute students across schools. This could indicate that at the time of reform the schools were relatively similar, since their classes had for a long time consisted of similar student.

[TABLE 8]

There is, however, a systematic pattern to the change in which schools different types of students attended before and after reform. Table 8 reveals some of the changes in composition in downtown high schools. Columns (1)–(5) show the proportion of high-ability downtown students at each downtown school in Bergen before and after reform. Column (6) reports the proportion at any downtown school, while Column (7) report the proportion of high-ability downtown students at private schools. The p-value of a two-proportion z-test for differences in proportions is reported in the last row. The table shows that high-ability downtown students moved between downtown high schools because of reform; specifically, they moved from Tanks and Bjørgvin to Katten and BHG. The table also shows that the reform did not influence the decision of high-ability downtown students of whether to attend downtown public schools or private schools.

All downtown high schools are not equally attractive to students. Since the reform induced more high-ability downtown students to attend a particular high school, the ATT could reflect a change in school quality, rather than simply peer effects. The school effects observed cannot explain all of the effect identified. The proportion of high ability downtown students at BHG and Katten increased by 18 percentage points as shown in Table 8. Even if the effect of attending BHG or Katten is large, a potential school effect can only explain a small proportion the identified effect of the reform.

8 Summary and Conclusion

There are many studies in the peer effect literature that rely on naturally occurring variation in peer characteristics to estimate peer effects. Commonly used natural experiments are school vouchers, desegregation schemes, or school assignment lotteries. This study used a school choice reform process in Bergen, Norway to investigate the effects of changes in peer characteristics at high schools for high-ability students.

A change from a catchment area-based intake system to a performance-based intake system, or the reverse, will have different consequences for different types of students. This study focused on a group—the high-ability downtown students in Bergen—for which the reform primarily resulted in a considerable shock in peer characteristics at high school. The ATT of the reform was identified by comparing the change in high school outcomes of this group of students before and after reform to comparable students in other cities. The analysis showed that this group of students attended high schools where peer students' average middle school GPA increased by 0.65 of one SD after reform, as against comparable groups in other cities.

The results showed that exam scores of downtown high-ability students in Bergen increased significantly due to the reform. Since the reform meant that this group of students attended high school with higher-ability peers, it is tempting to draw the conclusion from this finding that high-ability students gain from attending high school with other high-ability students. Secondary findings urge caution about the effect of the school choice reform on high-ability students, since the results suggested that middle school students adjust rapidly to the new high school intake system. Using a DDD identification strategy that aimed to account for middle school students' adjustment to school choice reform, we found some positive effects on exams taken at the end of high school.

Implementing a performance-based intake system is one way of creating a tracked system where similar students attend school together based on an achievement measure. Detailed policy recommendations regarding intake systems require more in-depth analysis on the total effects of the intake reform. The present study's results suggest that reform had a largely positive effect on students at all ability levels, although it may be more challenging to understand the underlying mechanisms that caused this effect in other groups.

References

Abdulkadiroğlu, A., Angrist, J., & Pathak, P. 2014. The elite illusion: achievement effects at Boston and New York exam schools. *Econometrica*, 82 (1), pp.137–196.

Angelucci, M., De Giorgi, G., Rangel, M. A., and Rasul, I. 2010. Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics*, 94 (3), pp.197–221.

Angrist, J. D. 2014. The perils of peer effects. *Labour Economics*, 30, pp. 98–108.

Angrist, J.D. & Lang, K. 2004. Does school integration generate peer effects? Evidence from Boston's Metco Program. *American Economic Review*, 94 (5), pp.1613–1634.

Betts, J. R., and Zau, A. 2004. Peer groups and academic achievement: Panel evidence from administrative data. Unpublished manuscript.

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

Bobonis, G. J., and Finan, F. 2009. Neighborhood peer effects in secondary school enrollment decisions. The Review of Economics and Statistics 91 (4), pp.695–716.

Bonesronning, H. 2008. Peer group effects in education production: Is it about congestion? The Journal of Socio-Economics 37 (1), pp.328–342.

Bonesronning, H., and Haraldsvik, M. 2014. Peer effects on student achievement: Does the education level of your classmates parents matter? Working paper.

Booij, A. S., Leuven, E., & Oosterbeek, H. 2017. Ability peer effects in university: Evidence from a randomized experiment. *The Review of Economic Studies*, 84 (2), pp.547–578.

Brugard, K. H. 2013. Does school choice improve student performance? Working paper.

Buchmueller, T. C., DiNardo, J., and Valletta, R. G. 2011. The effect of an employer health insurance mandate on health insurance coverage and the demand for labor: Evidence from Hawaii. *American Economic Journal: Economic Policy*, 3 (4), pp.25–51.

Carrell, S. E., Malmstrom, F. V., and West, J. E. 2008. *Peer effects in academic cheating. Journal of Human Resources*, 43 (1), pp.173–207.

Carrell, S. E., Sacerdote, B. I., and West, J. E. 2013. From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica*, 81 (3), pp.855–882.

Chetty, R., Hendren, N., & Katz, L.F. 2016. The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. *The American Economic Review*, 106 (4), pp.855–902.

Clark, D. 2010. Selective schools and academic achievement. *The BE Journal of Economic Analysis & Policy*, 10 (1).

Dahl, G. B., Løken, K. V., and Mogstad, M. 2014. Peer effects in program participation. *The American Economic Review*, 104 (7), pp.2049–2074.

DeSimone, J. 2009. Fraternity membership and drinking behavior. *Economic Inquiry*, 47 (2), pp.337–350.

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

Duncan, G. J., Boisjoly, J., Kremer, M., Levy, D. M., and Eccles, J. 2005. Peer effects in drug use and sex among college students. *Journal of Abnormal Child Psychology,* 33 (3), pp.375–385.

Hanushek, E. A., Kain, J. F., Markman, J. M., and Rivkin, S. G. 2003. Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18 (5), pp.527–544.

Haraldsvik, M. 2014. Does performance-based admission incentivize students? Working paper.

Hoxby, C. 2000. Peer effects in the classroom: Learning from gender and race variation. NBER Working Paper No. 7867.

Jackson, C.K. 2012. Single-sex schools, student achievement, and course selection: evidence from rule-based student assignments in Trinidad and Tobago. *Journal of Public Economics*, 96 (1), pp.173–187.

Jonsson, J. O., and Mood, C. 2008. Choice by contrast in Swedish schools: How peers' achievement affects educational choice. *Social forces*, 87 (2), pp.741–765.

Kling, J.R., Liebman, J.B., & Katz, L.F. 2007. Experimental analysis of neighborhood effects. *Econometrica*, 75 (1), pp.83–119.

Kling, J.R., Ludwig, J., & Katz, L.F. 2005. Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment. *The Quarterly Journal of Economics*, 120 (1), pp.87–130.

Lalive, R., and Cattaneo, M. A. 2009. Social interactions and schooling decisions. *The Review of Economics and Statistics*, 91 (3), pp.457–477.

Lavy, V., Paserman, M. D., and Schlosser, A. 2012. Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal*, 122 (559), pp.208–237.

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., & Sanbonmatsu, L. 2013. Long-term neighborhood effects on low-income families: evidence from moving to opportunity. *American Economic Review,* 103 (3), pp.226–231.

Machin, S., and Salvanes, K. G. 2010. Valuing school quality via a school choice reform. IZA Discussion Paper.

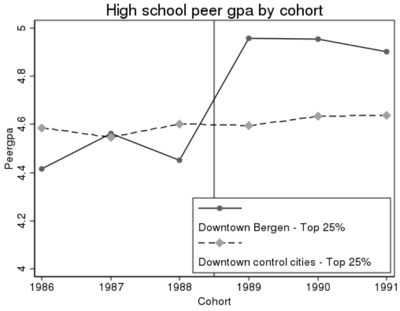
Manski, C. F. 1993. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60 (3), pp.531–542.

Rao, G. 2013. Familiarity does not breed contempt: Diversity, discrimination and generosity in Delhi schools. Working paper.

Sacerdote, B. 2011. Peer effects in education: How might they work, how big are they and how much do we know thus far? Handbook of the Economics of Education, 3, pp.249–277.

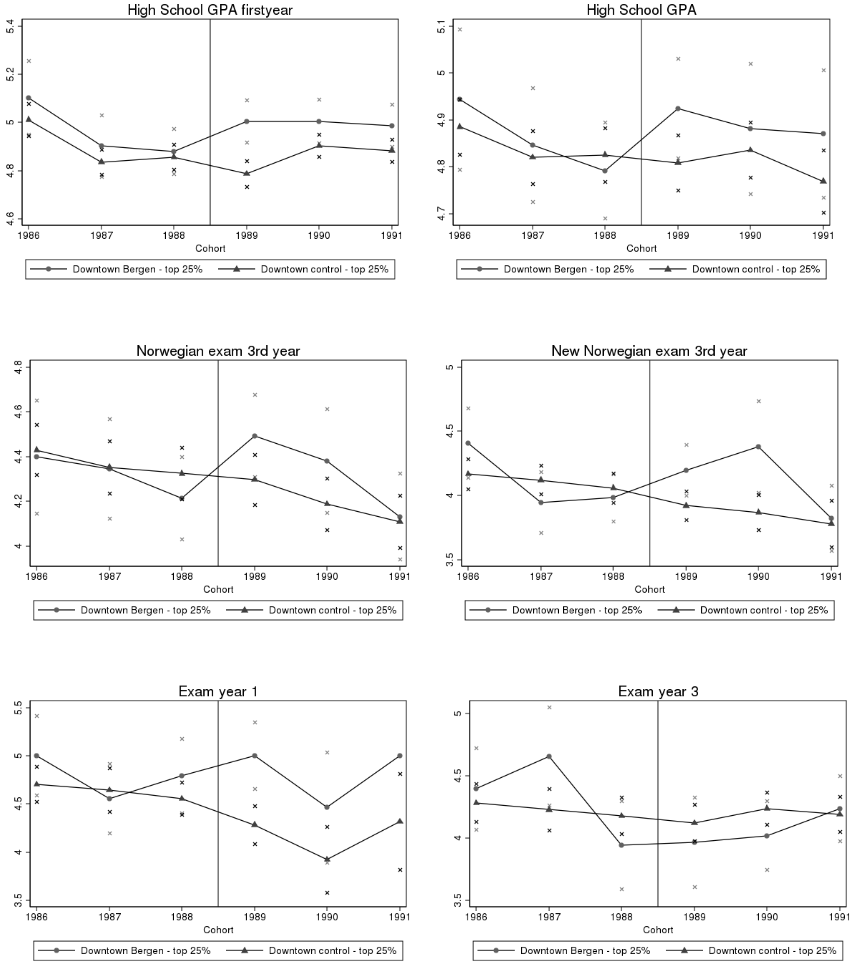
Wilson, J. 2007. Peer effects and cigarette use among college students. *Atlantic Economic Journal*, 35 (2), pp.233–247.

Figure 1



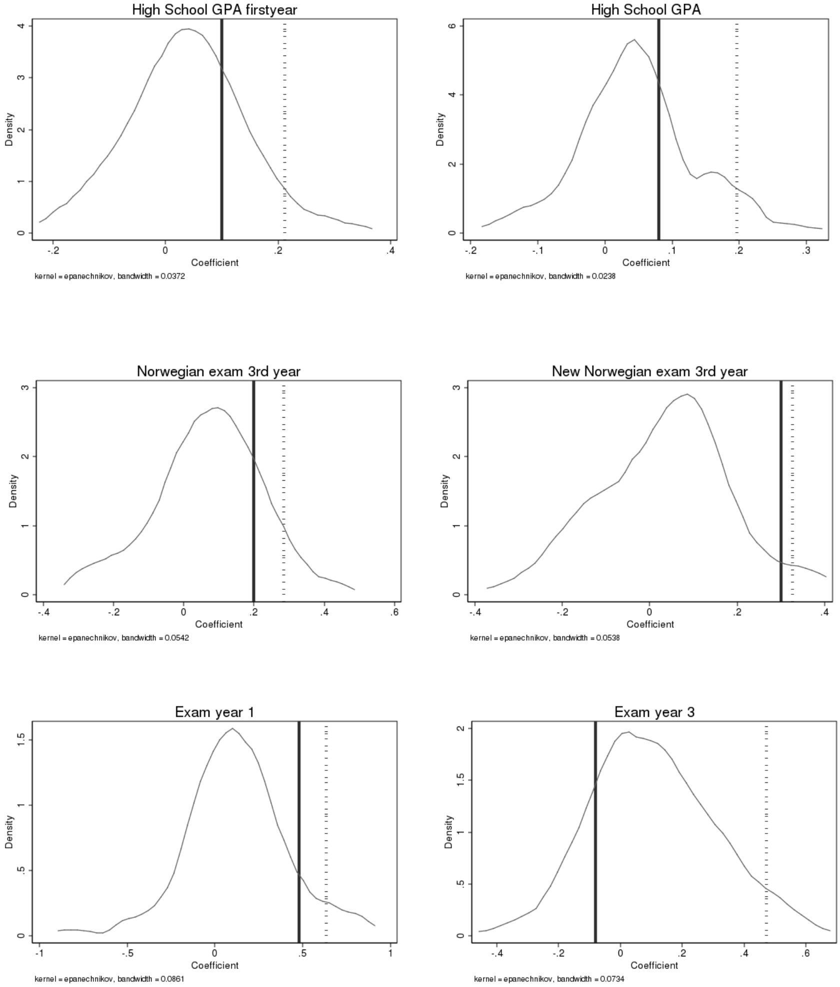
Notes: Figure shows peer gpa by treatment and control groups across cohorts. Cohorts born in 1989 finish middle school in the spring of 2005 and are the first that apply to high school after the school choice reform in Bergen. The treatment group consists of students that attended middle school downtown Bergen and are ranked among the top 25 % at middle school of their cohort in Bergen. The control group consists of students that attended middle school in Kristiansand, Stavanger, Trondheim or Drammen and are ranked among the top 25 % of their cohort.

Figure 2: Outcome trends DD



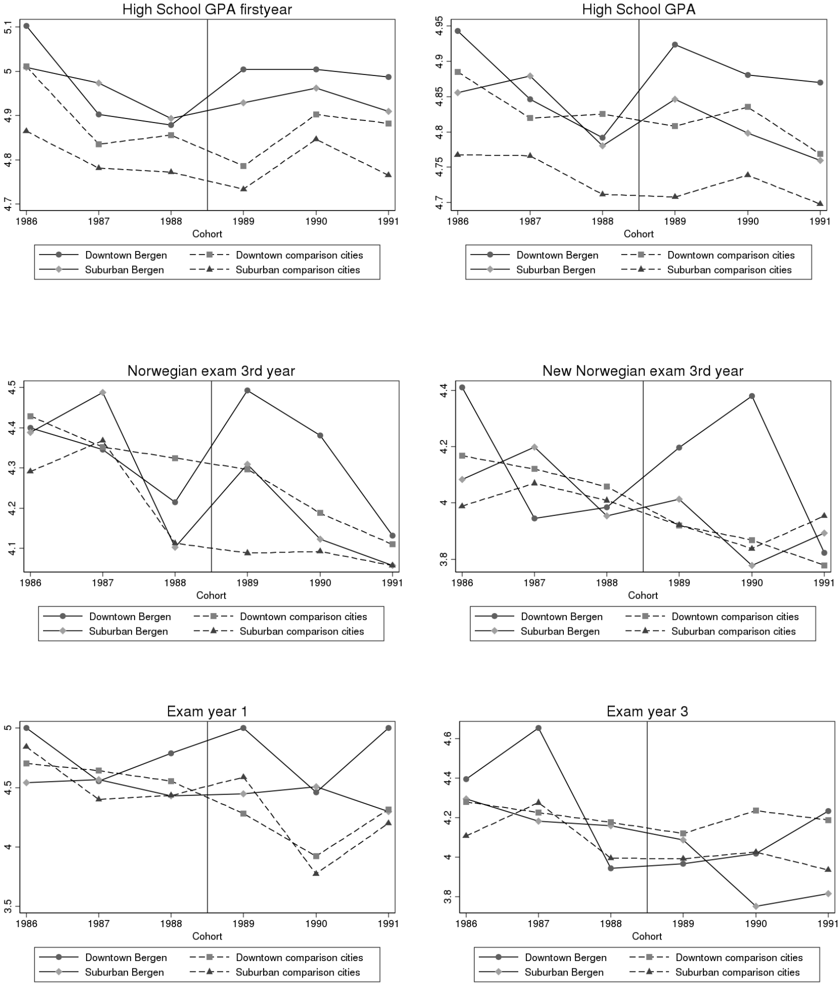
Notes: Figures show outcomes by groups over cohorts. 95 % confidence intervals of means are shown.

Figure 3: Permutation test



Notes: Figures show distribution of coefficents from estimating the effect of placebo reforms. We have assumed that the policy changed happened for each of the 20 largest cities in Norway, and estimated DD coefficient for the top 25 % in each of these cities as well as for the 3 other ability groups of students for each city. This gives a total of 80 coefficents. Dotted line represents the 95 percentile in the distribution of coefficients, while the full line is the estimate for Bergen.

Figure 4: Outcome trends DDD



Notes: Figures show outcomes by groups over cohorts. Only top 25 % of students included. 95 % confidence intervals of means are shown.

Table 1: Descriptive

	Treated		Diff: Treated - control		
	Pre reform		Pre reform	Post reform	
	Mean	SD	Mean	Mean	N
	(1)	(2)	(3)	(4)	(5)
Panel a) - Covariates					
Female	0.59	0.49	-0.08	-0.07	1869
Mother years of education	15.09	2.58	0.91	0.50	1736
Father years of education	15.83	2.68	0.93	0.42	1699
Mother earnings	100353	63944	5801	4492	1818
Father earnings	200699	136088	9561	-4503	1778
Middle School GPA	5.21	0.21	0.08	0.15	1869
Panel b) - Peer characteristic					
Peergpa	4.48	0.14	-0.10	0.31	1869
Panel c) - Outcomes					
Firstear GPA	4.94	0.47	0.05	0.14	1844
High School GPA	4.85	0.46	0.01	0.09	1656
Absence hours	32.23	30.11	-6.53	-7.73	1606
Absence days	14.19	11.60	-3.80	-1.50	1607
Select basic math year 1	0.05	0.21	-0.05	-0.01	1201
Norwegian exam year 3	4.31	0.82	-0.06	0.14	1734
Nynorsk exam year 3	4.09	0.86	-0.03	0.32	1342
Exam score year 1	4.76	0.81	0.13	0.62	420
Exam score year 2	4.49	0.99	0.02	-0.04	1040
Exam score year 3	4.25	1.19	0.02	-0.09	1432
# observations treatment			177	225	
# observations control			694	773	

Notes: Panel a) show descriptive statistics of covariates. Panel b) show endogenous variable peergpa. Peergpa are average middle school GPA of students at highs school. Panel c) show descriptive statistics of high school outcomes. Column (1) show pre reform means of the treated, Column (2) show the standard deviation of the treated. Columns (3) and (4) show difference in means between treatment and control before and after the reform. Treatment consists of students that attended middle school in the downtwon area of Bergen are ranked in the top 25 % of their cohort in Bergen. Control group consists of students that attended middle school in Kristiansand, Stavanger, Trondheim or Drammen and are ranked among the top 25 % of their cohort.

Table 2: Results

Dependent variable: High school outcomes

	(1)	(2)	(3)	(4)
Peergpa	0.42***	0.42***	0.43***	1869
	(0.09)	(0.07)	(0.07)	
GPA firstyear	0.09	0.09*	0.10*	1844
	(0.06)	(0.05)	(0.05)	
HS GPA	0.08	0.07	0.08	1656
	(0.06)	(0.06)	(0.05)	
Hours absent	-1.20	-1.03	-1.07	1606
	(4.38)	(4.33)	(3.97)	
Days absent	2.29	2.78	3.04	1607
	(2.66)	(2.32)	(2.24)	
Select basic math	0.04	0.05	0.04	1201
	(0.08)	(0.07)	(0.07)	
Norwegian exam year 3	0.20	0.20	0.20*	1734
	(0.12)	(0.12)	(0.11)	
Nynorsk exam year 3	0.34**	0.31**	0.30**	1342
	(0.14)	(0.14)	(0.12)	
Exam year 1	0.45*	0.46**	0.48***	420
	(0.24)	(0.21)	(0.18)	
Exam year 2	-0.00	-0.07	-0.02	1040
	(0.15)	(0.14)	(0.14)	
Exam year 3	-0.12	-0.10	-0.08	1432
	(0.20)	(0.19)	(0.18)	
Spesification				
Middle school dummies		x	x	
Background characteristics			x	

Notes: Table show coefficient estimates of β_3 in model 1 in colums (1)-(3) and sample size in Column (4). Sample consist of high ability downtown students in Bergen and control cities. Background characteristics are parents earnings, years of education and gender. Cohorts born 1986-1991. Standard errors in parentheses and are clustered at high school - year level. * p<0.10, ** p<0.05, *** p<0.01

Table 3: DD estimates for all groups in Bergen

Dependent variable: High school outcomes

	Downtown students				*Suburban students*			
Ability level	*L*	*M-L*	*M-H*	*H*	*L*	*M-L*	*M-H*	*H*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peergpa	0.03	0.13*	0.24***	0.43***	-0.03	0.04	0.11*	0.23***
	(0.09)	(0.08)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)	(0.07)
Firstyear GPA	0.17	0.21***	0.07	0.10*	0.16*	0.22***	0.15***	0.01
	(0.12)	(0.07)	(0.08)	(0.05)	(0.08)	(0.05)	(0.04)	(0.03)
High school GPA	0.22**	0.13**	0.13**	0.08	0.18***	0.20***	0.11***	0.02
	(0.09)	(0.06)	(0.07)	(0.05)	(0.06)	(0.04)	(0.04)	(0.04)
Hours absent	-7.41	-17.70**	-6.11	-1.07	2.36	-4.60	-2.61	1.71
	(11.18)	(7.39)	(6.43)	(3.97)	(7.42)	(4.82)	(3.38)	(3.28)
Days absent	2.32	2.94	2.03	3.04	4.27*	1.64	-0.77	-0.56
	(3.51)	(3.03)	(2.00)	(2.24)	(2.18)	(1.58)	(1.28)	(1.32)
Select basic math	0.02	-0.07	0.10	0.04	-0.07	-0.01	-0.09*	0.02
	(0.09)	(0.08)	(0.08)	(0.07)	(0.05)	(0.05)	(0.05)	(0.03)
Norwegian exam year 3	0.16	0.28**	0.08	0.20*	0.09	0.12*	0.13*	0.03
	(0.15)	(0.11)	(0.13)	(0.11)	(0.09)	(0.07)	(0.08)	(0.08)
Nynorsk exam year 3	0.03	-0.17	0.03	0.30**	0.02	0.13*	0.09	-0.03
	(0.18)	(0.11)	(0.11)	(0.12)	(0.09)	(0.07)	(0.10)	(0.08)
Exam year 1	0.24	0.64*	0.10	0.48***	0.24	0.28	0.23	0.17
	(0.39)	(0.33)	(0.20)	(0.18)	(0.21)	(0.18)	(0.16)	(0.15)
Exam year 2	-0.00	0.20	0.22	-0.02	0.09	0.10	0.09	0.15
	(0.18)	(0.18)	(0.18)	(0.14)	(0.12)	(0.12)	(0.12)	(0.11)
Exam year 3	0.01	0.11	-0.07	-0.08	0.22**	0.05	0.05	-0.20*
	(0.22)	(0.14)	(0.16)	(0.18)	(0.11)	(0.10)	(0.09)	(0.10)
Middle school dummies	x	x	x	x	x	x	x	x
Background characteristics	x	x	x	x	x	x	x	x
Observations	1957	1911	1859	1869	3525	3475	3332	3219

Notes: Table show coefficient estimates of β_3 in equation 1 for all groups. Column (1) compares the change in outcomes of low-ability downtown students from before to after the reform to low ability downtown students in control cities. Column (2) compares the change in outcomes of medium low ability downtown students in Bergen to medium low ability students in control citites. Observations refer to the number of students registered starting academic track. Cohorts born 1986-1991. Standard errors are shown in parentheses and are clustered at high school - year level. * p<0.10, ** p<0.05, *** p<0.01

Table 4: Placebo

High school outcomes			
	Placebo reforms		
	2003	2004	
	(1)	(2)	(3)
Peergpa	0.08	-0.10	871
	(0.09)	(0.09)	
GPA firstyear	-0.03	-0.07	847
	(0.07)	(0.07)	
HS GPA	-0.05	-0.08	814
	(0.06)	(0.07)	
Hours absent	5.06	1.46	796
	(7.25)	(5.29)	
Days absent	4.25	-1.14	796
	(2.86)	(2.98)	
Select basic math	0.01	0.00	871
	(0.05)	(0.05)	
Norwegian exam year 3	-0.05	-0.08	814
	(0.14)	(0.15)	
Nynorsk exam year 3	-0.36*	-0.12	797
	(0.20)	(0.17)	
Exam year 1	-0.21	0.28	271
	(0.23)	(0.23)	
Exam year 2	0.28	0.16	514
	(0.18)	(0.17)	
Exam year 3	-0.14	-0.48*	633
	(0.28)	(0.24)	
Spesification			
Background chars	x	x	
Middle school dummies	x	x	

Notes: Table show coefficient estimates of β_3 in equation 1. That is, a regression of future outcomes on reform dummy, treatment status dummy and interaction. In columns (1)-(2) the sample consists of pre reform cohorts. Column (1) sets the reform for the school year starting 2003 while Column (2) sets the reform for the school year starting in 2004. Column (3) show the sample size. Cohorts born 1986-1991. Sample includes high ability downtown students in Bergen and control cities. Standard errors are shown in parentheses and are clustered at high school - year level. * p<0.10, ** p<0.05, *** p<0.01

Table 5: Robustness

Dependent variable. High school outcomes						
			Control group		Keeping cohorts	
	Baseline	Top 33%	21 cities	1988-1989	1986-1989	1986-1990
	(1)	(2)	(3)	(4)	(5)	(6)
Peergpa	0.43***	0.40***	0.44***	0.53***	0.46***	0.45***
	(0.07)	(0.07)	(0.06)	(0.11)	(0.10)	(0.08)
GPA firstyear	0.10*	0.09	0.06	0.22***	0.17**	0.11*
	(0.05)	(0.06)	(0.05)	(0.08)	(0.07)	(0.06)
HS GPA	0.08	0.07	0.07	0.17**	0.10	0.07
	(0.05)	(0.06)	(0.05)	(0.08)	(0.07)	(0.06)
Hours absent	-1.07	0.71	0.33	-6.76	-5.56	-3.12
	(3.97)	(3.84)	(4.90)	(4.45)	(4.55)	(3.99)
Days absent	3.04	3.59*	3.06	3.04	2.59	1.90
	(2.24)	(2.00)	(2.09)	(3.26)	(2.69)	(2.37)
Select basic math	-0.02	-0.02	-0.02	0.05*	0.01	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Norwegian exam year 3	0.20*	0.14	0.15	0.33**	0.24*	0.25**
	(0.11)	(0.11)	(0.10)	(0.16)	(0.13)	(0.11)
Norwegian II exam year 3	0.30**	0.18	0.22*	0.45***	0.29*	0.36***
	(0.12)	(0.13)	(0.13)	(0.14)	(0.16)	(0.13)
Exam year 1	0.48***	0.43**	0.15	0.37*	0.53***	0.48**
	(0.18)	(0.17)	(0.16)	(0.19)	(0.19)	(0.19)
Exam year 2	-0.02	-0.03	0.07	-0.03	0.04	-0.10
	(0.14)	(0.15)	(0.12)	(0.19)	(0.17)	(0.15)
Exam year 3	-0.08	-0.24	-0.06	0.28	-0.09	-0.16
	(0.18)	(0.17)	(0.15)	(0.22)	(0.17)	(0.17)
Background char.	x	x	x	x	x	x
Middle school dummies	x	x	x	x	x	x
High school dummies						
Observations	1869	2420	12666	651	1201	1544

Notes: Table show coefficient estimates of β_3 in equation 1. That is, a regression of future outcomes on reform dummy, treatment status dummy and interaction. Column (2) keeps students ranked in the top 33 % of students within year and city. Column (3) includes students scoring in the top 25 % in more cities in Norway in the control. Column (4)-(6) only keeps students belonging to the cohorts indicated in the table header. Column 8 includes a spesification with high school dummies. Oslo not incuded. Cohorts born 1986-1991. Standard errors are shown in parentheses and are clustered at high school - year level. * p<0.10, ** p<0.05, *** p<0.01

Table 6: Selection and dynamic response

	Years of education		Earnings			
	Mother	Father	Mother	Father	Female	MS GPA
	(1)	(2)	(3)	(4)	(5)	(6)
Panel a)						
Short (α_5)	-0.24	-0.39	12679.12	-528.81	-0.02	0.08*
	(0.35)	(0.38)	(10350.45)	(18665.42)	(0.06)	(0.05)
Middle(α_6)	-0.37	-0.19	-13840.84	-22632.64	-0.07	0.03
	(0.62)	(0.46)	(9097.84)	(23835.71)	(0.06)	(0.03)
Long(α_7)	-0.67**	-0.87**	-5590.74	-12197.42	0.09	0.11***
	(0.33)	(0.40)	(11620.11)	(18273.99)	(0.08)	(0.03)
N	1736	1699	1818	1778	1869	1869

Panel b)		Norwegian	Nynorsk	Exam	GPA	HS GPA
	Peergpa	exam year 3	exam year 3	year 3	firstyear	
	(1)	(2)	(3)	(4)	(5)	(6)
Short (α_5)	0.46***	0.24*	0.27	-0.10	0.17**	0.10
	(0.10)	(0.13)	(0.17)	(0.18)	(0.07)	(0.07)
Middle(α_6)	0.43***	0.25	0.53***	-0.23	0.05	0.03
	(0.11)	(0.16)	(0.12)	(0.20)	(0.06)	(0.06)
Long(α_7)	0.38***	0.11	0.07	0.09	0.08	0.13
	(0.10)	(0.15)	(0.15)	(0.21)	(0.06)	(0.09)
N	1869	1734	1342	1432	1844	1656

Notes: Table show coefficient estimates of α_5, α_6 and, α_7 in model 2. Panel a show estimates when the dependent variable is predetermined background variables and middle school GPA, while panel b show estimates when dependent are high school outcomes. Middle school dummies included in all regressions, and covariates included in regressions in panel b. Sample includes high ability downtown students in Bergen and control cities. Cohorts born 1986-1991. Standard errors are shown in parentheses and are clustered at high school - year level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 7: Results II

Dependent variable: High school outcomes	DDD	
	(1)	(2)
Peergpa	0.19***	1869
	(0.06)	
GPA firstyear	0.09	1844
	(0.05)	
HS GPA	0.07	1656
	(0.06)	
Hours absent	-2.81	1606
	(-5.15)	
Days absent	3.38	1607
	(2.24)	
Select basic math	0.02	1201
	(0.07)	
Norwegian exam year 3	0.16	1734
	(0.12)	
Nynorsk exam year 3	0.34**	1342
	(0.14)	
Exam year 1	0.31	420
	(0.25)	
Exam year 2	-0.18	1040
	(0.16)	
Exam year 3	0.11	1432
	(0.19)	
Spesification		
Middle school dummies	x	
Background characteristics	x	

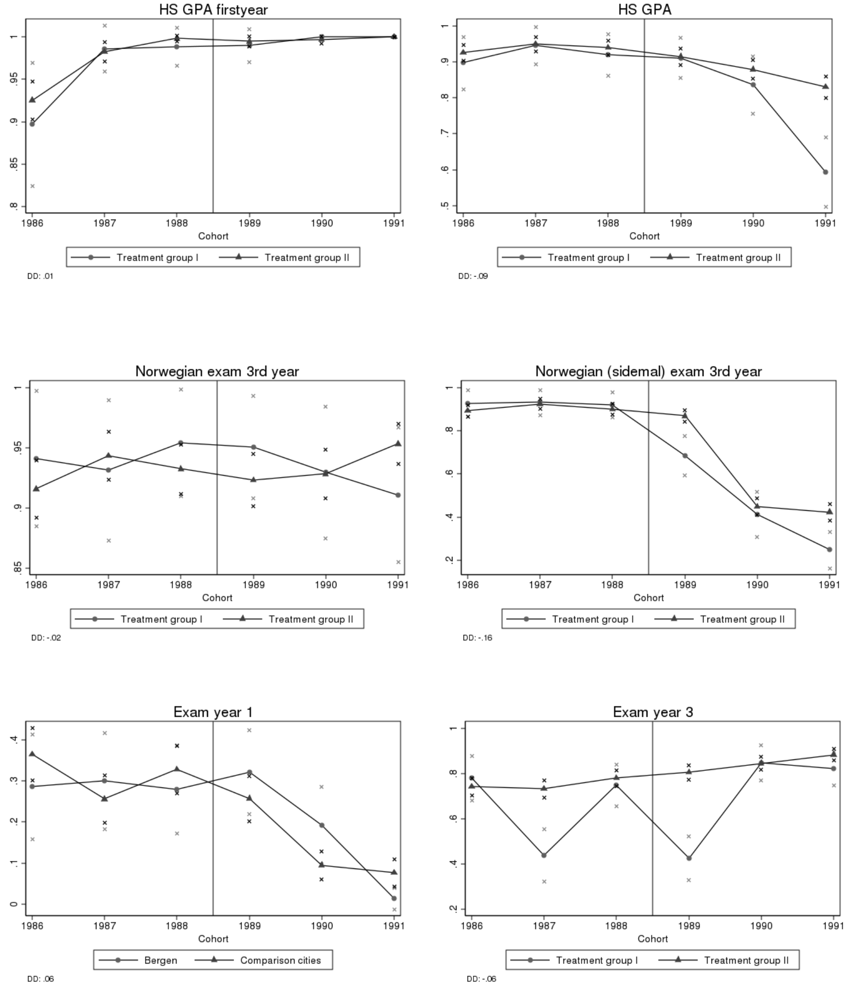
Notes: Table show coefficient estimates of γ_6 in Model 3 in Colum (1). Column (2) show sample size. Background characteristics are parents earnings, years of education and gender. Cohorts born 1986-1991. Standard errors are shown in parentheses and are clustered at high school - year level. * p<0.10, ** p<0.05, *** p<0.01

Table 8: Proportion of high ability downtown students attending school

	Katten	BGH	Langhaugen	Tanks	Bjørgvin	Downtown	Private
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Pre	0.23	0.12	0.25	0.26	0.04	0.90	0.05
Post	0.32	0.21	0.31	0.08	0.00	0.93	0.04
Diff	0.09	0.09	0.06	-0.18	-0.04	0.03	-0.01
P-value	0.04	0.03	0.20	0.00	0.01	0.37	0.97
Observations	402	402	402	402	402	402	402

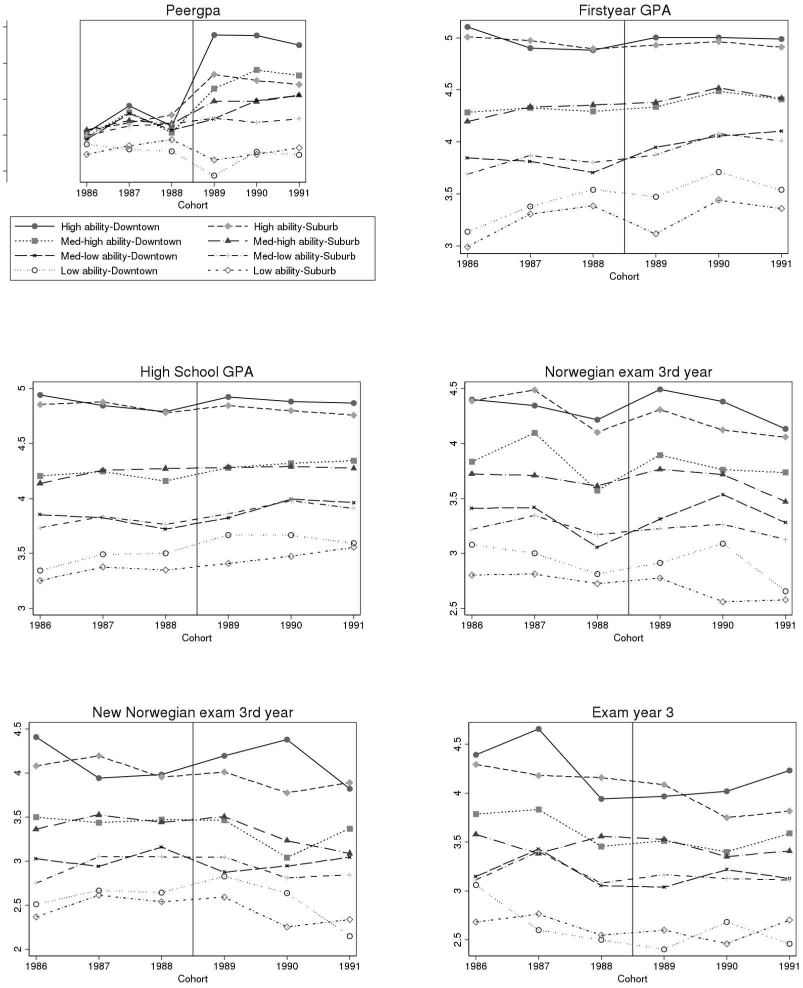
Notes: Table show proportion of high ability students attending each downtown high school in Bergen before and after reform in columns 1-5. Column 6 report the proportion at all downtown schools, while Column 7 report the proportion of high ability downtown students at private schools. P-values from a two-tailed test of proportions are reported.

Figure A.1: Attrition rate trends



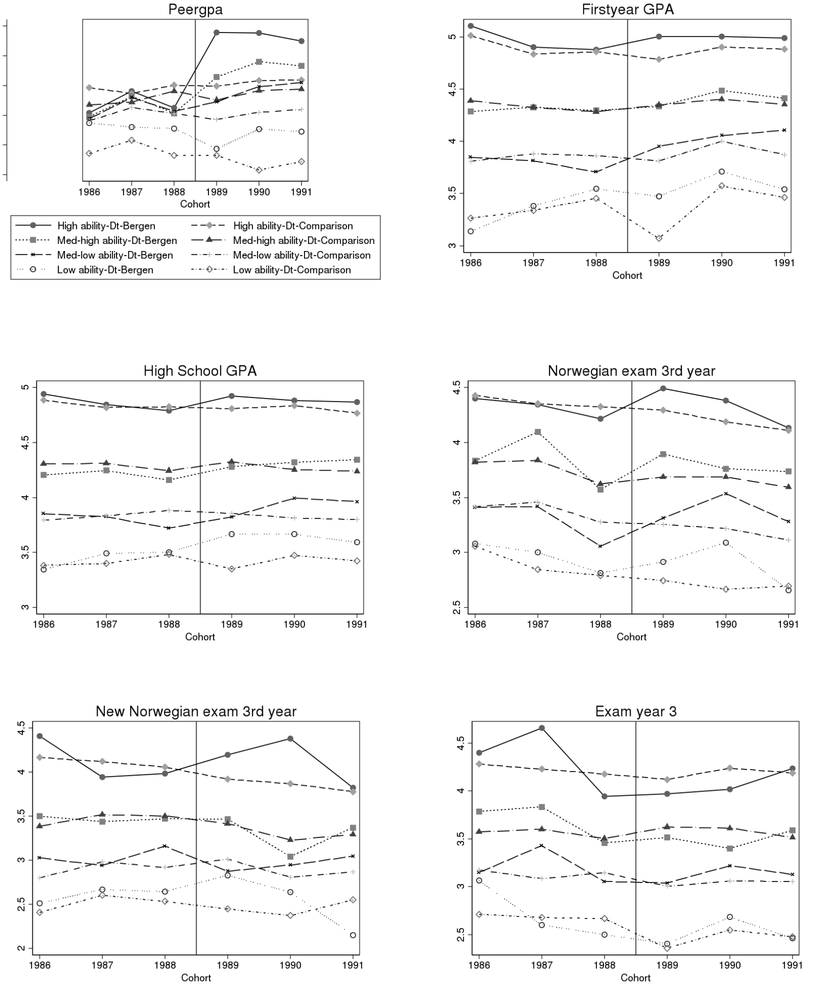
Notes: Figures show rates for which we observe test score outcomes over time. Difference in difference estimates are shown in the corner of each figure. Only significant difference in trends are detected for HS GPA at 10 % significance level. The jump in level of HS GPA firstyear from cohort 1986 is caused by a lack of administrative grades from the first year for this cohort.

Figure A.2: Descriptive figures - Bergen downtown and Bergen suburban students



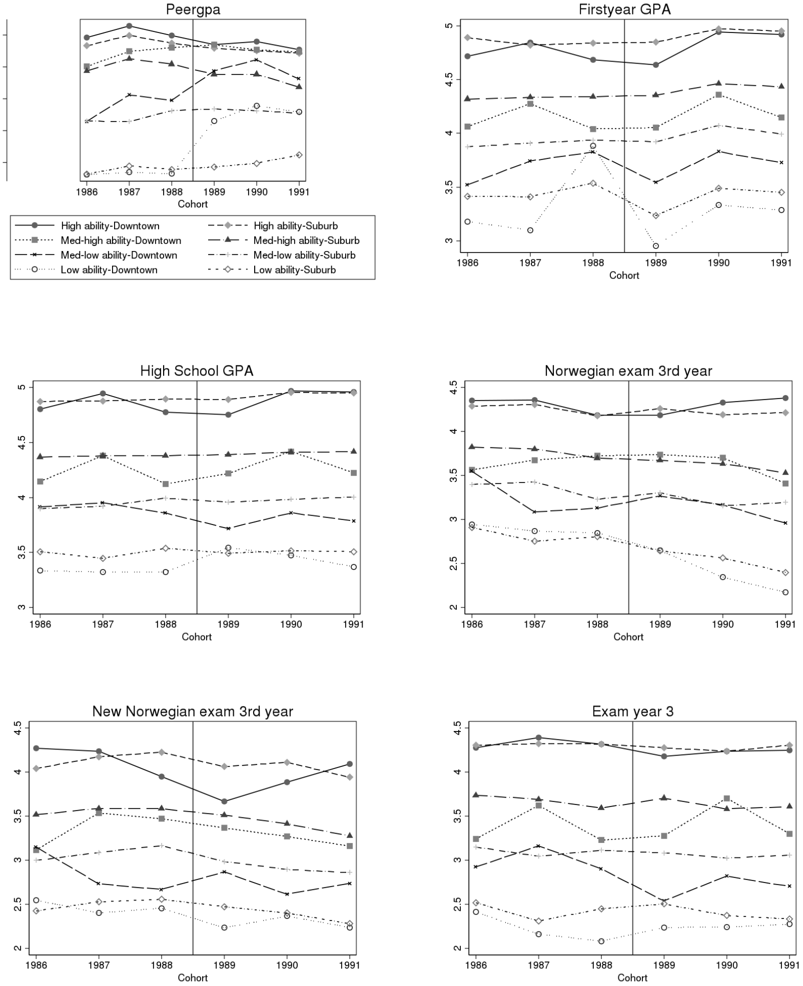
Notes: Figures show mean outcomes by cohort and group. Legends indicated in figure a.

Figure A.3: Descriptive figures - Bergen downtown and control downtown students



Notes: Figures show mean outcomes by cohort and group. Legends indicated in figure a.

Figure A.4: Descriptive figures - Oslo downtown and Oslo suburban students



Notes: Figures show mean outcomes by cohort and group. Legends indicated in figure a.

Table A.1: Estimation - Clustering on different levels

Dependent variable. High school outcomes	No	Middle school	High school	City	City 21	MS*year	HS*year
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Peergpa	0.42***	0.42***	0.43***	0.42***	0.44***	0.42***	0.43***
	(0.02)	(0.03)	(0.10)	(0.01)	(0.03)	(0.05)	(0.07)
GPA firstyear	0.10**	0.10**	0.10	0.10*	0.06***	0.10**	0.10*
	(0.05)	(0.05)	(0.06)	(0.03)	(0.02)	(0.04)	(0.05)
HS GPA	0.08	0.08**	0.08	0.08*	0.07***	0.08**	0.08
	(0.06)	(0.03)	(0.06)	(0.04)	(0.02)	(0.04)	(0.05)
Hours absent	-1.07	-1.07	-1.07	-1.07	0.33	-1.07	-1.07
	(4.31)	(3.04)	(3.30)	(3.23)	(1.23)	(2.92)	(3.97)
Days absent	3.04	3.04**	3.04	3.04***	3.06***	3.04*	3.04
	(2.12)	(1.36)	(1.93)	(0.48)	(0.49)	(1.67)	(2.24)
Select basic math	0.04	0.04	0.04	0.04	0.08***	0.04*	0.04
	(0.04)	(0.05)	(0.08)	(0.03)	(0.01)	(0.02)	(0.07)
Nor exam year 3	0.20*	0.20**	0.20***	0.20**	0.15***	0.20**	0.20*
	(0.10)	(0.09)	(0.07)	(0.06)	(0.03)	(0.07)	(0.11)
Norwegian II exam year 3	0.30**	0.30***	0.30***	0.30***	0.22***	0.30**	0.30**
	(0.12)	(0.08)	(0.07)	(0.03)	(0.03)	(0.11)	(0.12)
Exam year 1	0.48**	0.48***	0.48***	0.48**	0.15***	0.48***	0.48***
	(0.20)	(0.14)	(0.13)	(0.14)	(0.05)	(0.15)	(0.18)
Exam year 2	-0.02	-0.02	-0.02	-0.02	0.07*	-0.02	-0.02
	(0.16)	(0.09)	(0.17)	(0.15)	(0.04)	(0.15)	(0.14)
Exam year 3	-0.08	-0.08	-0.08	-0.08	-0.06**	-0.08	-0.08
	(0.14)	(0.08)	(0.11)	(0.11)	(0.03)	(0.18)	(0.18)
Spesification							
Middle school dummies	x	x	x	x	x	x	x
Background chars	x	x	x	x	x	x	x

Notes: Table show coefficient estimates of β_3 in model 1. Treatment group are high ability downtown students in Bergen, while control group are high ability students in control cities Trondheim, Savanger, Kristiansandand Drammen. Standard errors in parentheses. Cohorts born 1986-1991. Level of clustering indicated in column headers. City means clustering at city/muncipal level, while city 21 is clustering on city level when expanding to include 21 cities. MS*year is clustering on middle school - year, and HS*year is clustering on high school - year. * p<0.10, ** p<0.05, *** p<0.01

Table A.2: School quality

	Downtown students				Suburban students			
Ability level:	Low	Med-low	Med-high	High	Low	Med-low	Med-high	High
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Downtown								
Prop. pre	0.10	0.09	0.09	0.12	0.16	0.17	0.13	0.13
Prop. post	0.07	0.10	0.11	0.15	0.05	0.12	0.16	0.23
Diff	-0.03	0.01	0.02	0.02	-0.11	-0.05	0.03	0.10
P-value	0.01	0.57	0.05	0.11	0.00	0.00	0.02	0.00
Private								
Prop. pre	0.01	0.02	0.03	0.03	0.11	0.19	0.29	0.33
Prop. post	0.08	0.05	0.02	0.02	0.24	0.21	0.18	0.20
Diff	0.08	0.03	-0.01	-0.01	0.13	0.01	-0.10	-0.13
P-value	0.00	0.05	0.33	0.44	0.00	0.67	0.00	0.00
Katten								
Prop. pre	0.13	0.11	0.11	0.14	0.13	0.13	0.11	0.15
Prop. post	0.02	0.01	0.07	0.29	0.03	0.04	0.10	0.44
Diff	-0.10	-0.10	-0.03	0.15	-0.09	-0.09	-0.01	0.29
P-value	0.00	0.00	0.17	0.00	0.00	0.00	0.61	0.00
BHG								
Prop. pre	0.08	0.07	0.07	0.08	0.20	0.18	0.18	0.14
Prop. post	0.00	0.04	0.07	0.18	0.00	0.03	0.24	0.44
Diff	-0.08	-0.03	0.00	0.09	-0.20	-0.15	0.07	0.30
P-value	0.00	0.17	0.89	0.00	0.00	0.00	0.06	0.00
Bjorgvin								
Prop. pre	0.17	0.11	0.09	0.04	0.21	0.18	0.11	0.10
Prop. post	0.26	0.12	0.05	0.00	0.19	0.22	0.10	0.06
Diff	0.09	0.01	-0.04	-0.04	-0.02	0.04	-0.01	-0.05
P-value	0.02	0.77	0.13	0.00	0.59	0.26	0.86	0.07
Langhaugen								
Prop. pre	0.08	0.11	0.11	0.15	0.19	0.14	0.11	0.10
Prop. post	0.02	0.11	0.15	0.19	0.02	0.12	0.22	0.18
Diff	-0.06	-0.01	0.04	0.04	-0.17	-0.02	0.11	0.07
P-value	0.00	0.83	0.09	0.23	0.00	0.46	0.00	0.01
Tanks								
Prop. pre	0.06	0.07	0.09	0.17	0.11	0.21	0.15	0.14
Prop. post	0.06	0.21	0.22	0.07	0.02	0.20	0.14	0.08
Diff	-0.00	0.14	0.13	-0.10	-0.08	-0.02	-0.01	-0.06
P-value	0.90	0.00	0.00	0.00	0.00	0.59	0.78	0.03

Notes: Table show proportion students at each downtown school in Bergen before and after the school choice reform in column 1-5. Also show the proportion at all downtown schools, and the proportion students at private schools. P-value of a two-proportion z-test for differences in proportions are reported.

Table A.3: DDD placebo

	Years of education		Earnings			
	Mother	Father	Mother	Father	Female	MS GPA
$\hat{\gamma}_7$	-0.65**	-0.59*	-2326.65	635.34	0.01	0.04
	-0.33	-0.34	-8593.76	-16737.4	-0.07	-0.03
N	4760	4674	4974	4894	5081	5081

Notes: Table show estimates of γ_6 in model 3. Dependent variable are predetermined background variables and middle school GPA. Middle school dummies included in all regressions. Cohorts born 1986-1991. Standard errors are shown in parentheses and are clustered at high school - year level. * p<0.10, ** p<0.05, *** p<0.01

A Universal Childcare Expansion, Quality, Starting Age, and School Performance

Leroy Andersland[†]

This version: 03 August, 2017

Abstract

In the first decade of the 2000s, the proportion of children aged one or two attending formal childcare in Norway more than doubled, from 38% to 79%. There was an especially large increase in public funding to childcare centers starting in 2003, which accelerated this attendance growth. The consequences of this expansion on children`s outcomes remain largely unknown. This paper study the effects of attending childcare on school performance by using the fact that the childcare expansion was greater in municipalities that had low pre-reform childcare coverage. The results do not indicate any average effect of the childcare expansion on test scores at age 10. Dividing the municipalities into groups by childcare quality as measured by pre-reform observables, the results show a positive effect on school performance in municipalities with high pre-reform quality and a negative effect in municipalities with low pre-reform quality. Further analyses suggest that not only quality differences between municipalities but also the age of entering formal childcare explain the findings.

Keywords: Public Policy, Institutions, Childcare, Difference-in-Difference

JEL codes: D10, D02, J13, H31

[†] Department of Economics, University of Bergen, 5020 Bergen, Norway, email; lan004@uib.no

1 Introduction

There is an established consensus that high-quality targeted programs can have a positive impact on children's later life outcomes (Almond & Currie, 2010; Ruhm & Waldfogel, 2012; Heckman & Mosso, 2014). Meanwhile, there are still discussions in the literature on the consequences of public subsidized childcare programs (see examples of negative, no, and positive effects in Baker, Gruber, & Milligan, 2008; Gupta & Simonsen, 2010; Havnes & Mogstad, 2011, respectively). The main argument put forward for these differing findings is that children attending large-scale subsidized childcare programs have materially different profiles from the children in targeted programs and that the effect of attending childcare can vary for different types of children. For example, the consequence of attending childcare can vary for children from low or high socioeconomic backgrounds, by child age, by child personality, and based on the quality of the alternative modes of care as compared to the quality of the childcare program. Other important discussions in the literature are whether childcare has the potential to influence cognitive or non-cognitive outcomes or both, and whether the effects of childcare in early life persist through adolescence and into adulthood.[1]

This paper contributes to this debate by examining the consequences of a large expansion in public subsidized childcare following a 2003 reform in Norway. Outcomes of children in municipalities that on average experienced a large increase in childcare capacity is compared to the outcomes of children in municipalities that on average experienced a smaller increase in childcare capacity. Since the expansion mostly influenced one- and two-year-olds in some municipalities and three- to five-year-olds in others, this paper contributes evidence on the effects of childcare on children of different age groups. Unlike the studies using differential expansion across districts at the time of reform (Havnes & Mogstad, 2011; Felfe,

[1] Evidence from neurobiological studies shows that the level of stress hormones in children in childcare varies by childcare quality, time in care, age, and temperament (Gunnar & Donzella, 2001; Geoffroy et al., 2006; Lupien et al., 2009).

Nollenberger, & Rodŕıguez-Planas, 2015; Cornelissen et al., 2015; Felfe & Lalive, 2015), we contribute by using a slightly different estimation strategy, relying on expansions being stronger in municipalities with low pre-reform coverage. The advantage of this method is that endogeneity is less of a concern, since pre-determined characteristics of the municipality are used to identify high- and low-expansion municipalities. The outcomes available are national test scores performed at age 10. In addition,childcare quality is examined, since rich data on quality measures are available for childcare institutions. Recent reports on the importance of childcare quality for child outcomes have motivated this focus (e.g., Walters, 2015; Araujo et al., 2015). This paper contribute to this literature by examining the direct effect of a rapid expansion of coverage on child outcomes in municipalities with different pre-reform quality at childcare centers.[2]

The empirical strategy builds on studies that use differential expansions of public programs across districts to evaluate the impact of such programs. An early example of this procedure can be found in Duflo (2001, 2004), whose articles explore the consequences of a major primary school construction program in Indonesia in the 1970s. A notable feature of the expansion program was the intention to construct more schools in areas with relatively low school coverage. The results show clear positive impacts on years of completed education and adult earnings for those who were more exposed to the school construction program. Since the program led to a doubling of the number of schools in six years, Duflo (2001) also examines what happened to school quality. Focusing on the pupil-teacher rate, this indicated that that quality declined between the pre- and post-reform periods, but that there was no differential decrease between high-program and low-program areas.

[2] It is important to note that observable inputs are not randomly assigned to municipalities. If the effect of an expansion varies between municipalities with different observables, that may actually be due to differences in unobservable dimensions, different mechanisms in different types of municipalities, or groups of children being differently affected in different types of municipalities. Section 6 attempts to explore some of these issues.

The literature on the consequences of expansions of large-scale public childcare programs reports mixed effects on child outcomes. Analysis of such programs is most often performed in countries with both an advanced public childcare sector and high-quality administrative data. Canada, Denmark, Norway, and Germany are examples of such countries.[3] Baker, Gruber, & Milligan (2008) published an early paper on the analysis of universal childcare. A program to increase childcare availability was implemented in one Canadian province, Quebec, but not in the rest of the country. The coverage rate for children aged four and under in Quebec increased in the reform period, which lasted nine years, from about 43% to about 67%. The article compares the child outcomes in Quebec with other Canadian jurisdictions, finding negative effects on short-term behavioral and health outcomes for both children, who were assessed soon after leaving the program before age five, and for the program, since the effects were estimated not long after it was implemented.[4] The paper does not find that quality of care decreased by looking at indicators of staff qualifications (age, proportion full-time, proportion with some secondary education) in Quebec and the rest of Canada.

Havnes & Mogstad (2011) examined the effect of a large-scale childcare coverage increase mainly for children aged three to six in Norway in the 1970s. Studying a reform that took place several decades earlier allowed them to look at long-term outcomes, and they find positive effects on adult labor market participation and years of education. With regards to quality, they conclude that, if anything, childcare quality fell because child-teacher and child-staff rates increased more in those municipalities that expanded their coverage to a greater

[3] Evidence on the effect of a relatively large-scale program (compared to Perry Preschool or the Abecedarian Project) from the US is found in studies based on Head Start (Currie & Thomas, 1995; Ludwig & Miller, 2007; Garces, Thomas, & Currie, 2002; Deming, 2009; Bitler, Hoynes, & Domina, 2014; Walters, 2015). Gormley et al. (2005), Wong et al. (2008), and Fitzpatrick (2008) report on public preschool programs in different states.

[4] Lefebvre, Merrigan, & Verstraete (2008) similarly find negative effects using the same reform on preschool children's Peabody Picture Vocabulary Test scores. Baker, Gruber, & Milligan (2015) show that negative effects persist into young adulthood. Other studies pointing to negative effects of childcare (not based on expansions) are Bernal & Keane (2010) and Herbst & Tekin (2010).

degree. By looking more closely at the distributional effects of the same reform on earnings, Havnes & Mogstad (2015) show a positive effect driven by those at the lower end of the outcome distribution. Other evidence from Norway mainly that relies on causal identification methods different than capacity expansions are found in Drange & Havnes (2015), Drange & Telle (2015), and Drange, Havnes, & Sandsør (2016). They find either no effect or a positive effect of attending childcare.

Gupta & Simonsen (2010) examined a policy in Denmark that guaranteed access to childcare for certain cohorts in some municipalities. The study does not find any average effect of center-based childcare in Denmark on a range of measures of non-cognitive outcomes. They do find a negative effect of family daycare for boys with mothers with vocational-track education. Felfe, Nollenberger, & Rodŕıguez-Planas (2015) studied an expansion in childcare capacity in Spain and find positive effects, driven by girls and disadvantaged children, on reading and math scores at age 15. Cornelissen et al. (2017) looked at a capacity expansion aimed at 3-6 year olds in Germany, employing a marginal treatment effects (MTE) framework to show positive effects on children from disadvantaged backgrounds, which also manifests in the MTE analysis as a "reverse selection on gains": those that are less likely to participate are the same people that have the most to gain from childcare. Felfe & Lalive (2014) studied the effect of universal childcare before age three in Germany and find a positive effect for children from low socioeconomic backgrounds. Smaller increases in capacity appear to be more beneficial than larger increases.

The political motivation behind large-scale public childcare expansions and targeted childcare interventions differ to some degree, since the former is driven more by increasing the parental labor supply and the latter more directly at child development. Thus, ex ante, what to expect about the effects on children of large-scale expansions should be less clear. In this study, we find no average effect of a childcare expansion in Norway on 5th grade test scores. Furthermore, we do not find any consistent evidence that childcare quality is affected by the expansion. Looking at the effect of childcare in municipalities with a high level of quality in childcare institutions before reform, as measured by a pre-reform indicator, the consequence of the expansion is positive on school performance. Negative effects are found in districts with low childcare quality before reform and are mainly driven by children with low socioeconomic status (SES). Further analyses were not able to show that childcare quality or maternal labor force participation were affected differently in municipalities with different levels of pre-reform quality. However, the children influenced by the expansion were of different ages in the two types of municipalities. In low-quality municipalities, the expansion largely affected one- and two-year-olds, while in the high-quality municipalities, the expansion primarily involved three to five-year-olds. This is an indication that childcare might be negative for the youngest children, especially in an environment of lower-quality childcare and positive for older children, especially in an environment of higher-quality childcare.

The study proceeds as follows. In section 2, background information on the institutional details of the reform and the childcare sector is provided. Section 3 describes the administrative data used and lays out the empirical strategy of difference-in-differences using pre-reform coverage rates as a predictor of childcare supply shocks. Section 4 presents the main results, while Section 5 focuses on childcare quality. Section 6 examines alternative mechanisms and Section 7 summarizes and concludes the paper.

2 Institutional details

The development of a public childcare sector in Norway is related to the increase in female labor market participation. In the mid-1960s, few mothers were active labor market participants, so there were relatively few childcare centers. As women's participation accelerated in the 1970s, there was a corresponding increase in childcare attendance. In the early stages of the development of a public childcare sector in Norway, the focus was on offering alternatives to older children aged three to six years. This has changed over time, with the labor market attachment by mothers of younger children increasing substantially.

The 1990s were subject to three reforms that had important consequences for the public childcare sector. In 1993, maternity leave was extended up until children turned one, and the process of including six-year-olds in the school system was finalized in 1997. As a result, most children attending public childcare were aged between one and five after 1997. In 1998, the Cash-for-Care (CFC) benefit was implemented, providing a substantial cash incentive to parents that did not send their one- or two-year-olds to childcare. The reform showed that a price increase reduces childcare attendance, and mainly increase parental care for the youngest children (Andersland & Nilsen, 2016).

By the beginning of the 2000s, the public childcare sector in Norway was already well developed. About 41% of children aged one or two and about 84% of children aged three to five attended some form of public childcare in 2002.

[FIGURE 1]

Figure 1 shows the development of the proportion of one- and two-year-olds registered in childcare in Norway; the rapid increase in the coverage rate in the 2000s is readily apparent. This expansion is associated with “The Childcare agreement,” a 2003 decision by the Storting, Norway’s legislature, that changed several laws to increase childcare capacity. The most important elements were the equal treatment of childcare centers, the implementation of guaranteed childcare for those who wished it, and the implementation of a nationwide maximum price for childcare.

Norway’s formal childcare system features both private and public operations. Both types of centers receive public subsidies and are subject to similar regulations. Before reform, however, municipalities differed in the amount of subsidy provided to private childcare centers. The reform meant that all types of childcare centers were to receive the same subsidy amount. Moreover, municipalities were obliged to guarantee a place in public childcare by 2005 for all children by 1st September of the year following their birth. The government made plans to create 40,000 new childcare slots by 2005.

Before reform, prices could vary substantially between municipalities. Beginning in 2004, the monthly maximum price was set at 2,750 NOK (≈US$340), with plans for eventual decreases. A survey conducted in 2002 reports that municipal childcare prices averaged by parental income groups ranged from 2,044 to 2,937 NOK (Eibak, 2002). The introduction of a fixed price would thus largely affect high-income households in practice, since they paid the highest prices for childcare before reform.

To implement a maximum price, equal treatment of childcare centers, and guaranteed childcare slots, Norway's total public expenditure for childcare more than doubled from 2002 to 2005.[5] In addition to increasing regular funds, the government established a discretionary fund to help municipalities that would face particularly daunting challenges in fulfilling the reform requirements (Aamodt, Moennesland, & Juell, 2005). This fund would direct extra resources to those municipalities that had higher prices for and lower subsidies to private childcare centers before reform and to those considered likely to be unable to guarantee childcare slots for all who needed them by 2005. The specific amount of funds directed to each municipality was calculated centrally. The net effect was that more funds were directed at municipalities with less-developed childcare sectors. The discretionary fund amounted to 10.5% of Norway's total public funding for childcare in 2005.

Formal childcare in Norway is centrally regulated through the Childcare Act ("Barenhageloven"), which provides a set of common rules for childcare in municipalities

[5] Source: National Budgets 2002–2007.

throughout the country. The maximum number of children in full-time care per pedagogical leader (a childcare position type that requires certified education) is nine for children below three and eighteen for children from three to six years old. To become a pedagogical leader, one must complete three years of college education in the preschool teacher program. Childcare centers can apply to the municipality for temporary exemptions if they are not able to meet the educational requirements for staff. There are both caregivers with formal education and caregivers without pedagogical education working in childcare. The stated norm is one caregiver for three children below three and one caregiver for six children three or older. Childcare centers are normally open during daytime working hours, from 7 am to 5 pm, Monday to Friday. Statistics on the use of different forms of care arrangements can be useful for understanding alternative forms of care for one- and two-year-olds. A survey conducted by Statistics Norway in 2002 and reported in Pettersen (2003) shows that 44% of children in that cohort are cared for primarily by parents, 33% attend formal childcare, 12% have informal care arrangements, 4% are cared for by relatives, and 7% have other care arrangements. These figures make clear that a large increase in one- and two-year-olds attending formal childcare will be draw primarily from parental care and other informal care arrangements.

3 Data and empirical strategy

Data on cohorts born from 1998 to 2004 are used in the analysis. These cohorts were chosen because they were affected by the childcare expansion and because they are now old enough to provide data on school performance at age 10.[6] Information on childcare attendance comes from two sources. The measure of individual-level childcare attendance comes from Norway's CFC database. The CFC benefit was implemented in 1998 to provide a cash

[6] Table A1 shows the structure of the data.

transfer to families that did not send one- or two-year-old children to formal childcare. Families applied to the welfare agency to receive the benefit, stating whether or not their child attended childcare. From 1999–2012, a family would be eligible for the benefit simply by having a one- or two-year-old child who did not attend formal childcare fulltime. The welfare agency controlled this information by collecting monthly information from municipalities. As a result, individual-level childcare measure for all CFC-eligible children from 1998 onward are available. Children are classified as attending childcare in a given month if they were registered as attending above 10 hours per week that month. From these data, we also construct a measure of the total number of hours in childcare before age three.

Municipality-level childcare attendance rates, or coverage rates, are taken from the KOSTRA ("Kommune-Stat-Rapportering") database. KOSTRA is a national reporting scheme used in Norway for the administration, evaluation, and comparison of municipalities. Childcare centers report their numbers at the end of each year to their municipality, after which municipalities report the number of children in childcare to Statistics Norway. This database is used to calculate municipality-level coverage rates and municipality-level childcare quality measures.[7,8]

The third important source of information is the database on national exams, which were introduced in 2004 in order to evaluate how schools succeeded in developing students' skills in math, reading, and English. Students take the tests at the 5th, 8th, and 9th grade levels. Since the students in our sample are still very young, we can only examine the effects on 5th grade test scores. Since 2008, the tests in math and English have been electronically corrected. Depending on the subject and year, scores are given on a scale from 0 to 30 or 50.

[7] Figure A1 compares municipality coverage rates calculated from CFC data with KOSTRA coverage rates by year and across the largest municipalities.

[8] KOSTRA coverage rates and quality numbers are reported for most municipalities from 2001 onward. Since we use numbers registered in the second year after the cohort birth year, the 1998 cohort is excluded from the analysis using KOSTRA numbers. To compensate for this, the robustness section shows the main results without the 1998 cohort.

In this analysis, the distribution of points is standardized with means of 0 and standard deviations (SDs) of 1 by subject and year.

Information is also available on which school students attended when taking the test, municipality of residence by age, and parents' earnings and years of education. Earnings are measured in basic amounts used in the national insurance scheme.[9]

The empirical strategy follows previous literature on universal childcare by using a large expansion in capacity that varies across districts. The first step is to identify a pre-reform indicator for the intensity of expansion; examples of similar strategies can be found in Duflo (2001) and Løken, Lundberg, & Riise (2017).

[TABLE 1]

Table 1 shows the results from a regression of municipality-level coverage expansion for one- and two-year-olds from 2002–2007 on pre-reform municipality characteristics. It reveals that the pre-reform coverage rate (for one- and two-year-olds measured in 2001) is a strong predictor of capacity expansion. The 2003 reform led to a larger increase in childcare capacity in municipalities with lower initial coverage rates. A municipality with 10 percentage points (pp.) higher pre-reform coverage had a 4.75 pp. lesser increase in the coverage rate. This result accords with the regulatory changes following the reform that led to more funds being directed at municipalities with less-developed childcare sectors. The pre-reform coverage rate is therefore used as the indicator of the capacity expansion and the main regression (Equation 1) is:

$$Y_{it} = \alpha_1 + \alpha_2 Short_t + \alpha_3 Long_t + \alpha_4 (PreCoverage_i \cdot Short_t) + \alpha_5 (PreCoverage_i \cdot Long_t) + \alpha_6 X_{it} + \epsilon_{it} \quad (1)$$

[9] One basic amount was 92,576 NOK in 2016 ≈ US$11,443.

The cohorts in the sample were born between 1998 and 2004. $Short_t$ is an indicator for cohorts born in 2001 or 2002 and only partly affected by the reform, while $Long_t$ is an indicator for cohorts born in 2003 and 2004 and more significantly affected by the reform. $Short_t$ will also be an indicator for children that are affected by the reform when they are older, while $Long_t$ is an indicator for children that are more affected by the reform when they are younger. $PreCoverage_i$ is the pre-reform municipal level childcare coverage rates, measured in 2001, for one- and two-year-olds. X_{it} is a set of control variables including gender, mother's age, father's age, immigration status, parents' labor participation before child birth, parental years of education before child birth, and municipality dummies. Y_{it} are average scores on national tests in math, English and reading in the 5th grade. To correct for intragroup correlation in error terms, standard errors are clustered at the municipality level.

The specification assumes a linear relationship between the pre-reform coverage rate and outcome variables. Following Løken, Lundberg, & Riise (2017), municipalities above the 90th and below the 10th percentiles in the pre-reform coverage rate distribution are dropped from the analysis.[10] Municipalities with very high or low pre-reform coverage rates may behave differently than other municipalities in response to the reform because of their extreme pre-reform coverage rates. We show the sensitivity of the results to this restriction in the robustness section.

The interpretation of the coefficient in front of the interaction terms in Equation 1, with test scores as outcomes, is how changes in test scores from before to after the reform depend on pre-reform coverage levels. The coefficient is interpreted as an intention-to-treat (ITT) effect, since it is the total effect on children in municipalities more heavily exposed to

[10] Figure A2 shows the coverage rate distribution with lines indicating the 10th and 90th percentiles. Figure A3 explores how pre-reform characteristics relate to pre-reform coverage rates in municipalities. The figure reveals that municipalities with higher pre-reform coverage rates on average had higher female employment, lower male employment, larger cohort sizes, more private childcare centers, lower proportions of preschool teachers, and lower adjusted costs per care hour.

the childcare expansion. For our estimate to have a causal interpretation, we must assume that without the childcare expansion, the time trends in test scores in municipalities with high and low pre-reform coverage rates would have been the same.

[FIGURE 2]

Figure 2a illustrates the empirical strategy by showing the change in coverage rates from 2002 to 2007, compared to pre-reform municipal-level coverage rates of for one- and two-year-olds in 2001. There is a clear relationship between pre-reform coverage rates and capacity increases. Municipalities with a relatively high pre-reform coverage level of close to 0.6 showed an average increase in coverage level of about 0.2, while municipalities with a relatively low pre-reform coverage level of nearly 0.2 had an average increase in coverage level of about 0.4. Figure 2b shows that there is no significant relationship between pre-reform coverage levels and changes in test scores, which is the first sign that there was no average effect of the reform on test scores.

4 Results

This section begins the exploration of the effects of the expansion on test scores in detail. Estimations are carried out on both on the full sample and, in the next section, on subgroups of municipalities in an attempt to see if the effect of the expansion varied across different types of municipalities. Lastly, we explore the mechanisms behind the estimated results. Table 2 shows the effect of the expansion for children using a sample containing all municipalities.

[TABLE 2]

The first two columns in Table 2 show the results when the dependent variable is average national exam test scores with and without individual-level control variables. The short run coefficient is insignificant, small, and negative, while the long run coefficient is insignificant, small, and positive. Including control variables changes the estimates only marginally. Thus, there are no conclusive signs that higher exposure to childcare expansion affected school performance at age 10. Columns 3–5 show the same estimations with individual controls by subjects: reading, English, and math, respectively. The only significant coefficient is in the short run effect for math. This may seem inconsistent with a positive effect of childcare since the expansion is much stronger in the long run. One explanation for finding significant effects in the short run but not the long run is that the effect of childcare is heterogeneous by age. The cohorts affected by the expansion in the short run estimate (born 2001–2002) were older when the reform began to take effect. This issue is discuss this in greater detail in Section 6.

Table 3 shows the robustness of the baseline results. The different specifications are indicated in the column headers.

[TABLE 3]

Column 1 repeats the baseline results from Table 2, while Column 2 excludes Norway's six largest cities from the estimation. The long run estimate changes sign, but both coefficients remain insignificant. Column 3 excludes the 25% smallest municipalities, with no significant effect on the estimates. Column 4 excludes municipalities with pre-reform coverage rates below the 15th or above the 85th percentile in the pre-coverage rate distribution. The short run coefficient appears to be somewhat sensitive to changing the pre-reform coverage cutoff rates, becoming negative and significant at the 5% level. Column 5 excludes the first cohort in the pre-reform period, while Column 6 excludes the first cohort in the post-reform period. Excluding cohorts has only a marginal impact on the size of the coefficients. Column 7 interacts predetermined municipal characteristics with cohort dummies to determine whether if municipalities with different observable characteristics demonstrate different trends. Observable characteristics used are female labor force participation, male labor force participation, and cohort size. Coefficients remain small and insignificant.

The estimates are generally robust to changes in the specifications. The only specification change that appeared to matter was changing the cutoff in the pre-reform coverage distribution, and even that change was observed only for the short run coefficient. This could be a sign that there are nonlinear effects but could also be an artifact of chance. In total, Table 3 show that the results are robust to a variety of specifications checks.

5 Childcare quality

Until this point, the focus has been on the average effect of the childcare expansion on child outcomes. This section has two main objectives. First, it explores how the expansion affects childcare quality. Second, it seeks to determine whether the effect of the expansion depends on pre-reform municipality-level childcare quality. To achieve these two objectives, good measures (or correlates) of municipality childcare quality must be obtained. The KOSTRA database provides a set of potential variables for this purpose. The quality measures available are "Children/staff rate," "Adjusted care hours/staff rate," "Proportion preschool teachers among pedagogical leaders," "Proportion preschool teachers among employees," "Cost per child," and "Cost per adjusted care hour."

The "Children/staff-rate" and "Adjusted care hours/staff-rate" measure how much exposure each child has to a caregiver (or group size) while in childcare. Staff is measured in person-year full-time equivalents; it's thus not sensitive to changes in the use of part-time staff. Adjusted care hours are the number of hours of childcare provided, adjusted for the age composition of the children, which is determined by multiplying the number of children below three by two and the number of children aged three by 1.5 and giving a factor of one to children aged three or older. In the time period we study, Statistics Norway has only calculated this measure for public childcare centers.

The general norms imply that the proportion of pedagogical leaders of employees is the same for personnel working with children aged one to two and children aged three to five. As a result, the proportion preschool teachers among employees should not be determined by the change in age composition. As mentioned, all pedagogical leaders are supposed to have preschool teacher education. "Proportion of preschool teachers among pedagogical leaders" should therefore not be affected by the change in age composition, even if childcare centers operate with different employment structures than those suggested by the norms. These proportions are calculated for all childcare centers within a municipality.

"Cost per child" and "Cost per adjusted care hour" are measures of how much municipalities spend on childcare. Since it costs more to keep younger children in childcare, "Cost per child" does depend on age composition. As with the adjusted group size measure, Statistics Norway has only calculated these measures for public childcare centers.

The selection of group size measures was motivated by the literature, which shows that class size matters (Krueger, 1999; Chetty et al., 2011). The municipality database does not include information on the experience of childcare employees, which has been shown to be an observable teacher characteristics that is a relevant correlate of teacher quality in the school literature (Rivkin, Hanushek, & Kain, 2005; Staiger & Rockoff, 2010), but it does include information on the education levels of employees in childcare centers. Since those measures are made at the municipality level, we argue that they do not necessarily reflect the quality of the individual childcare employees; rather, they reveal something about the overall quality of childcare in a given municipality. According to national regulations, pedagogical leaders are supposed to have a certificate in preschool education. A municipality that lacks a high proportion of preschool teachers among its pedagogical leaders suggests either that it has problems recruiting and retaining quality staff or that it is not strict in adhering to standards in childcare centers.

Three cautionary remarks are necessary. First, as Rivkin, Hanushek, & Kain (2005) note, there are large differences in teacher quality that are not easily captured by readily available observable characteristics. This suggests that our measures of municipality-level teacher education will capture only some quality differences across municipalities and centers. Second, even though these may be policy relevant variables, they do not measure the actual interactions in childcare centers as observed in Araujo et al. (2016). Lastly, observable inputs are not randomly assigned to municipalities. If the effect of the expansion varies between municipalities with different observables, that may actually be due to their being different on unobservable dimensions, that the mechanism is different in different types of municipalities, or that the groups of children affected are dissimilar in different types of municipalities. Section 6 offers an initial exploration of some of these issues.

5.1 Effect of expansion on observable inputs

With a large expansion in public subsidies to childcare centers it is not clear ex ante whether one should expect an increase or decrease in childcare quality. The expansion led to a doubling of the funding for childcare centers in just three years, so one could reasonably expect increased funding to enhance quality. However, a rapid increase in the number of children in childcare centers could also lead to lower quality by increasing group size and lowering the qualification and experience levels among the personnel. The literature reviewed in Section 1 suggests that if anything, quality normally falls with large-scale expansions.

To analyze the effect of childcare expansion on childcare quality, we estimate Equation 2, which has municipality-level quality measures as dependent variables:

$$\begin{gathered} Quality_{it} = \alpha_1 + \alpha_2 Short_t + \alpha_3 Long_t + \alpha_4 (PreCoverage_i \cdot Short_t) \\ + \alpha_5 (PreCoverage_i \cdot Long_t) + \alpha_6 X_{it} + \epsilon_{it} \quad (2) \end{gathered}$$

[TABLE 4]

Table 4 shows the results from these estimations. The dependent variable is average municipality level childcare quality from the year in which the children in each cohort turn two through the year in which they turn five. Information for the 1998 cohort is dropped, since we lack quality information for this cohort (data on quality measures only begins in 2001). Column 1 examines how age composition in childcare changes, Columns 2–3 examine the effect on measures of group size, Column 4–5 look at measures of the quality of staff, and Columns 6–7 examine how childcare costs are affected by the expansion.

Column 1 shows that children in municipalities with 10 pp. lower pre-reform coverage rates attended childcare centers with a 1.08 pp. higher proportion of children aged one or two in childcare in the short run, and a 2.21 pp. higher proportion of children aged one or two in childcare in the long run. This confirms the hypothesis that the proportion of children aged one or two in childcare centers was affected by the expansion; any analysis of how the expansion changes quality across time should take account of this reality.

Column 2 shows that the number of children per staff decreases in high-expansion municipalities. The long run estimates show a decrease of 0.17 (3.6% of the mean or 29% of the SD) children per caregiver in childcare in municipalities with a 10 pp. lower pre-reform coverage rate. Children are normally divided into groups by age, with a fixed number of adults responsible for each group. Even with the significant change brought on by the reform, children's exposure to adults may actually be unchanged, since the age composition in childcare changes.[11]

[11] The organization of childcare centers can be divided into "Avdelingsbarnehage" and "Basebarnehage". "Basebarnehage" allows for a more open organization that lets children roam between groups, although each child should still keep a main attachment to a specific group with a fixed number of adults (Vassenden et al., 2011). The more children are allowed to roam across age groups, the regular child pr. staff measure of childcare quality becomes important.

Column 3 shows no effect on adjusted care hours/staff for municipal childcare centers, indicating that there is no evidence that children's exposure to adults changes as a consequence of the expansion. Since private and public childcare centers are subject to the same regulations, it is likely that group size also remains unchanged in private childcare centers.

In addition to group size, employee education levels and amount of municipal spending on childcare may indicate how an expansion affects quality. Column 4 shows no change in the proportion of staff with preschool teacher education, while Column 5 similarly shows no significant increase in the total proportion of pedagogical leaders with preschool teacher education. Together, these are interpreted as meaning no average change in the education level of childcare center staff as a consequence of the reform. Column 6 shows that costs per child in childcare did increase as a consequence of the expansion, but Column 7 indicates that the increase is relatively smaller per age-adjusted care hour. Given that we do not find any evidence of a change in group size in public childcare centers or in employee education level, we are cautious about how to interpret this coefficient. It could mean that expansion leads to increased quality through other channels, but it could also mean that efficiency declines during a capacity buildup, since costs increase for the same number of age-adjusted care hours provided.

In sum, the results indicate that group size and education level among employees remain unchanged, while childcare costs increase to some extent. Since the increased costs are relatively small and may actually signal decreased efficiency instead of increased quality, we conclude that we are not able to reject the hypothesis that childcare quality is unchanged. Additional analysis of how changes in observable inputs relate to changes in child outcomes is provided in Section 6.

5.2 Effect on test scores of the child care expansion by municipality type

To uncover possible heterogeneity in effects across, municipalities are split into groups according to the proportion preschool teachers of employees and the proportion of preschool teachers among pedagogical leaders before reform. According to the norms and regulations, these proportions should stay constant across municipalities with different age compositions in childcare. Any observed variation is therefore more likely to be due to quality differences. These measures cover both private and public childcare centers. Table 5 shows the results of estimation carried out using subsamples.

[TABLE 5]

The table shows estimates of the short and long run coefficients from Equation 1, split into four panels. The dependent variables are average test score, reading score, English score and math score in Panels a), b), c), and d) respectively. In Columns 1–3, Equation 1 is estimated separately by dividing municipalities according to three quantiles in the distribution of pre-reform municipality-level proportion preschool teachers among employees. In Columns 4–6, the same approach is carried out using the proportion of preschool teachers among pedagogical leaders.

Since results are very similar across the two measures of quality, we choose to focus on the results in Columns 4–6. The long-run ITTs on average test scores show that the expansion affected test scores negatively among municipalities with the lowest proportion of preschool teachers among pedagogical leaders before reform, while expansion led to an increase in test scores in municipalities with the highest such proportion. The same is true when dividing municipalities according to proportion preschool teachers among employees. The long run effect shows that children in municipalities with 10 pp. lower pre-reform coverage rates increased test scores in high-quality municipalities by 0.041 SD, while it decreased by 0.028 SD in low-quality municipalities. This pattern is fairly consistent across different test scores, but it is most prominent in the long run estimates, once reform has had time to be implemented and exert greater influence. Table A2 shows the robustness of results when dividing municipalities according to the proportion of preschool teachers among pedagogical leaders, with average test scores as the outcome as shown in Panel A in Table 5. The table shows that estimates for high-quality municipalities are robust to different specifications, while the estimates for low-quality municipalities are somewhat sensitive to the exclusion of large municipalities and flexible trends.[12]

One possible reason for why these measures may capture childcare quality is that they describe how easy it is for municipalities to hire quality personnel in childcare. Even though pedagogical leaders are supposed to have certified preschool education, the proportion that actually has this certification varies between municipalities. The regulations acknowledge that it may be a challenge to hire qualified personnel: they therefore allow municipalities to apply for exemptions. However, observable inputs are not randomly assigned to municipalities. There are thus alternative explanations that are discussed in the next section.

6 Alternative mechanisms and heterogeneous effects

The previous section indicated that the reason for positive effects of childcare on children's test scores in municipalities with a high pre-reform proportions of preschool teachers and negative effects on children's test scores in municipalities with low pre-reform proportions of preschool teachers could be quality differences between childcare centers in different types of municipalities. Table 6 allows us to explore alternative explanations.

[Table 6]

Column 1–3 in Table 6 show estimations of long run coefficients in Equation 1 on subgroups of municipalities split according to the proportion of preschool teachers among pedagogical

[12] See Section 6 for a further discussion of this issue.

leaders in a municipality. Each column shows results for estimations on subsamples, with row headings indicating dependent variables. Panel A) shows results using municipality-level childcare quality measures as outcomes, while Panel B) has parental labor force participation as outcomes, and outcomes in Panel C) are childcare attendance measures.

Panel A) shows that the children/staff rate decreases significantly in low-quality municipalities as a consequence of the expansion, while high-quality municipalities experience a smaller decrease in that measure. At the same time, there are no significant changes in the adjusted care hours/staff rate in any of the municipality groups. These results are consistent with a change in age composition in childcare centers that are different in the different municipality types, but does not indicate that different developments in group size can explain the observed differences in results between municipality types, because we observe no differences in adjusted care hours per employee. This pattern repeats itself in the results on the effect of the expansion on childcare costs. While the cost per child increases significantly with expansion in low-quality municipalities, we are not able to reject the null hypothesis of no change in costs per adjusted care hour in either high- or low-quality municipalities. The change in education level appears to be unaffected in all municipality types, although there does appear to be an increase in the education level among pedagogical leaders in low-quality municipalities. However, this cannot explain the negative effect of the expansion in these municipalities since, if anything, an increase in the education level of staff should translate into a positive effect on test scores.[13] Children/staff-rate decreases significantly in low-quality municipalities as a consequence of the expansion. If children are free to roam between groups, this would indicate an increase in quality due to increased caregiver exposure, but since we observe a negative effect for this group of municipalities, this does not appear to be a quantitatively important explanation. In sum, group size,

[13] The results shown in Table A3 show that the only significant change between high- and low-quality municipalities on quality measures is on the children/staff rate.

education level, and costs do not evolve very differently in the different municipality types, at least not in the direction expected from the differences in test scores.

Effects on parental labor force participation are displayed in Panel B). Parental labor force participation is defined as a parent's earning above two basic amounts in the national insurance scheme in the second year after a child's birth (Havnes & Mogstad, 2011, 2015). Increased labor force participation could explain a positive effect of the childcare expansion by increasing household incomes (Løken, 2010; Black et al., 2012; Dahl and Lochner, 2012; Løken, Mogstad, & Wiswall, 2012). A positive effect of the expansion is found on maternal but not paternal labor force participation. Importantly, the point estimates do not suggest that household income changes differentially in different municipality types. This suggests that income effects cannot fully explain the differences in results across different types of municipalities.[14]

Lastly, Columns 1–3 in Panel C) show the effects of the expansion on childcare attendance in different types of municipalities. The regressions suggest that expansion leads to a strong increase in childcare attendance before age three in low-quality municipalities, while the same effect is not observed in high-quality municipalities. This is consistent with the results on municipality-level coverage rates, which show a strong increase in experienced coverage rate for one- and two-year-olds in low-quality municipalities, while there is a much smaller increase in high-quality municipalities. The largest increase in coverage rates for three- to five-year-olds are found in high-quality municipalities. These findings are also consistent with the patterns on unadjusted group size and costs measures. The expansion leads to increased childcare attendance mostly for children aged one or two in low-quality municipalities, while it mainly increases attendance for children aged three to five in high-quality municipalities. The main alternative explanation for the negative effect observed in

[14] Table A4 shows results for the full sample, and Table A3 shows results from tests of different development in parental labor force participation in high- and low-quality municipalities. No significant difference is found.

low-quality municipalities is therefore that the effect of childcare on young children is negative. The positive effects observed in high-quality municipalities are then explained by positive effects of childcare for older children. This might be an important explanation for the mixed findings in the literature, since papers more often find positive effects for three- to five-year-olds (e.g., Havnes & Mogstad, 2011) and no or negative effects for one- or two-year-olds (e.g., Fort, Ichino, & Zanella, 2016)

Table A5 show the robustness of results shown in Table 6 by using pre-reform coverage rates for children aged one to five to estimate ITT effects. Panel c) reveals that this pre-reform indicator is associated with a shift of both more and relatively older children into childcare for low-quality municipalities. With the childcare quality explanation being the only explanation, we should expect to see stronger effects of the same sign. Column 4 in Panel d) shows that the effect in the low-quality municipalities is no longer significantly different from zero. This is consistent with the explanation that there are negative effects of attending childcare for children aged one or two in low-quality municipalities, while this is not necessarily so for older children. This robustness therefore suggests that the effects originally found in Table 5 are likely to be partly explained by the heterogeneous effects of childcare by starting age.

Table A6 shows the results from a regression of a municipality's pre-reform characteristics on a dummy indicating whether the municipality has a high or low proportion of preschool teachers among pedagogical leaders (excluding the group of municipalities in the middle). Before reform, the proportion of children in childcare was lower in municipalities that expanded coverage for the oldest children the most. We also note that differences in quality measures are statistically insignificant or small in measures other than the education level of employees. The average difference is 0.21 pp. in the proportion of preschool teachers among pedagogical leaders between the two municipality types. The difference is noticeable, but indicates a very large return on better-educated childcare personnel if this is the only explanation behind the different effects. It therefore suggests that the municipalities differ according to unobserved measures, or that child age is an important explanation in the finding of different effects.

[TABLE 7]

It is possible to look at the heterogeneous effects of the expansion according to observables. Table 7 suggests that the positive effects found in municipalities that mostly expanded access to older children are driven by females and high-SES children, while the negative effects in municipalities that mostly expanded access to younger children are driven by females and low-SES children. The observed positive effects for females are consistent with other findings in the literature (Anderson, 2008; Havnes & Mogstad, 2011; Felfe, Nollenberger, & Rodŕıguez-Planas, 2015). The largest point estimate is the negative effect for low-SES children in municipalities that mostly expanded access to younger children.[15]

Not finding any positive effect for low-SES children is the most surprising item in Table 7. Differential effects by SES may indicate that the childcare quality to which each group is exposed could be different, if low-SES children are more likely to be exposed to lower quality childcare centers within municipalities.

7 Conclusion

Publicly subsidized childcare and targeted childcare programs differ in many respects. Previous research has shown mixed findings on the effect of large-scale public subsidized childcare programs on child outcomes. This study adds evidence to the literature by providing an analysis of a recent expansion in childcare capacity combined with new high-quality administrative data from Norway. Large expansions of universal childcare are costly, and

[15] Table A7 shows robustness for this finding. It confirms that the negative effect is driven by low-SES children, as the estimate is less sensitive to specification changes than when not restricting the sample based on child background characteristics.

analyzing the effects of these programs is important both for countries in which programs already have been implemented and for governments that are evaluating whether to implement similar public programs.

In contrast to earlier studies of the effects of public subsidized childcare expansions, we use a pre-reform indicator to identify municipalities with high or low expansions. In line with a Norwegian reform passed in 2003, we find that children in municipalities with low pre-reform coverage received increased access to childcare compared to those with higher pre-reform coverage. Using this pre-reform coverage measure as an indicator for childcare expansion, we do not find that the expansion has an average impact on test scores at age 10. The analysis proceeded by looking for heterogeneous effects by dividing municipalities according to observable inputs to childcare. Looking at the effect of the expansion among municipalities with a high level of childcare quality, as indicated by the proportion of preschool teachers among pedagogical leaders, we find positive and significant effects on child test scores of attending childcare. At the same time, we find a negative effect on children's test scores of the expansion in municipalities with the lowest proportions of preschool teachers among pedagogical leaders. Notably, the negative effects appear to be driven mainly by low-SES children. Further analysis was not able to show that childcare quality and maternal labor force participation were affected differently by pre-reform quality. However, the children influenced by the expansion were of different ages in the two types of municipalities. In low-quality municipalities, the expansion largely affected one- and two-year-olds, while the expansion mostly affected three- to five-year-olds in the high-quality municipalities. This is an indication that childcare might be negative for the youngest children, especially in an environment of lower-quality child care, and positive for older children, especially in an environment of higher-quality child care.

Given the significant increase in childcare coverage for the youngest children in recent years, it is important to know the relative significance of these two explanations. If the effect of attending childcare for young children is negative only in low-quality municipalities, then the proper policy implication will be to increase quality in these areas. More research is needed to understand the full effects on all one- and two-year-olds.

References

Aamodt, G., Moennesland, J., & Juell, E. 2005. *Finansiering av barnehagene.* Temanotat 2005/3. Bergen. Utdanningsforbundet.

Almond, D. & Currie, J. 2011. Human capital development before age five. *Handbook of Labor Economics*, 4, pp.1315–1486.

Andersland, L. & Nilsen, O.A. 2016. *Cash-for-care benefits and formal childcare attendance.* [Online]. http://folk.uib.no/secaa/Public/Trygd2014/cashcare200916.pdf. Accessed 30th May 2017.

Anderson, M.L. 2008. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103 (484), pp.1481-1495.

Araujo, M.C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. 2016. Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics*, 131 (3), pp.1415–1453.

Baker, M., & Gruber, J., & Milligan, K. 2008. Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy*, 116 (4), pp.709–745.

Baker, M., & Gruber, J., & Milligan, K. 2015. *Non-cognitive deficits and young adult outcomes: the long-run impacts of a universal child care program*. NBER Working Paper 21571. [Online]. http://www.nber.org/papers/w21571. Accessed 30th May 2017.

Barenhageloven, "Lov om barnehager", Norges lover (2005).

Bernal, R. & Keane, M.P. 2010. Quasi-structural estimation of a model of childcare choices and child cognitive ability production. *Journal of Econometrics*, 156 (1), pp.164–189.

Bitler, M.P., Hoynes, H.W., & Domina, T. 2014. *Experimental evidence on distributional effects of Head Start*. NBER Working Paper 20434. [Online]. http://www.nber.org/papers/w20434.pdf. Accessed 30th May 2017.

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

Borge, L.E., Johannesen, A.B., & Tovmo, P. 2010. Barnehager i inntektssytemet for kommunene. Senter for økonomisk forskning.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. 2011. How does your kindergarten classroom affect your earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126 (4), pp.1593–1660.

Cornelissen, T.., & Raute, A., Schönberg,, U., & Dustmann, Christian. 2017. Who benefits from universal child care? Estimating marginal returns to early child care attendance. *Forthcoming: Journal of Political Economy*.

Currie, J. & Thomas, D. Does Head Start make a difference? 1995. *The American Economic Review*, 85 (3), pp.341–364.

Dahl, G. B. & Lochner, L. 2012. *The impact of family income on child achievement: Evidence from the earned income tax credit*. The American Economic Review, 102 (5), pp.1927-1956

Deming, D. 2009. Early childhood intervention and life-cycle skill development: evidence from Head Start. *American Economic Journal: Applied Economics*, 1 (3), pp.111–134.

Drange, N. & T. Havnes. 2015. Child care before age two and the development of language and numeracy: Evidence from a lottery. IZA discussion paper 8904 [Online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2582539. Accessed 2nd August 2017

Drange, N. & Telle, K. 2015. Promoting integration of immigrants: effects of free child care on child enrollment and parental employment. *Labour Economics*, 34 (C), pp.26–38.

Drange, N., Havnes, T., & Sandsør, A.M.J. 2016. Kindergarten for all: long-run effects of a universal intervention. *Economics of Education Review*, 53, pp.164–181.

Duflo, E. 2001. Schooling and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *The American Economic Review,* 91 (4), pp.795–813.

Duflo, E. 2004. The medium run effects of educational expansion: evidence from a large school construction program in Indonesia. *Journal of Development Economics*, 74 (1), pp.163–197.

Eibak, E.E. *Undersøking om foreldrebetaling I barnehager, august 2002*. 2002. Statistics Norway Notes 2002/29. Oslo: Statistics Norway.

Felfe, C. & Lalive, R. 2014. *Does early child care help or hurt children's development?* IZA Discussion Paper 8484. [Online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2505346. Accessed 30th May 2017.

Felfe, C., Nollenberger, N., & Rodŕıguez-Planas, N. 2015. Can't buy mommy's love? Universal childcare and children's long-term cognitive development. *Journal of Population Economics*, 28 (2), pp.393–422.

Fitzpatrick, M.D. 2008. Starting school at four: the effect of universal pre-kindergarten on children's academic achievement. *The BE Journal of Economic Analysis & Policy*, 8 (1), pp.1–40.

Fort, M., Ichino, A., & Zanella, G. 2016. *Cognitive and non-cognitive costs of daycare 0-2 for girls*. Quaderni working paper DSE 1056. [Online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737370. Accessed 30th May 2017.

Garces, E., Thomas, D., & Currie, Janet. 2002. Longer-term effects of Head Start. *The American Economic Review*, 92 (4), pp.999–1012.

Geoffroy,M.C., Côté, S.M., Parent, S. & Séguin, J. R. Daycare attendance, stress, and mental health. *The Canadian Journal of Psychiatry*, 2006, (51), pp.607-615

Gormley, W.T., Gayer, T., Phillips, D., & Dawson, B. 2005. The effects of universal pre-K on cognitive development. *Developmental Psychology*, 41 (6), pp.872–884.

Gunnar, M.R. & Donzella, B. Social regulation of the cortisol levels in early human development. *Psychoneuroendocrinology*, 27 (1). pp.199–220.

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

Havnes, T. & Mogstad, M. 2011. No child left behind: subsidized child care and children's long-run outcomes. *American Economic Journal: Economic Policy*, 3 (2), pp.97–129.

Havnes, T. & Mogstad, M. 2015. Is universal child care leveling the playing field? *Journal of Public Economics*, 127, pp.100–114.

Heckman, J.J. & Mosso, S. 2014. The economics of human development and social mobility. *Annual Review of Economics, Annual Reviews,* 6 (1), pp.689–733.

Herbst, C.M. & Tekin, E. 2010. Child care subsidies and child development. *Economics of Education Review*, 29 (4), pp.618–638.

Krueger, A.B. 1999. Experimental estimates of education production functions. 1999. *The Quarterly Journal of Economics*, 114 (2), pp.497–532.

Lefebvre, P., Merrigan, P., & Verstraete, M. 2008. *Childcare policy and cognitive outcomes of children: results from a large scale quasi-experiment on universal childcare in Canada*. CIRPEE Working Paper 08-23. [Online]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1267335. Accessed 30th May 2017.

Ludwig, J. & Miller, D.L. 2007. Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 122 (1), pp.159–208.

Lupien, S.J., McEwen, B.S., Gunnar, M.R., & Heim, Christine. 2009. Effects of stress throughout the lifespan on the brain, behaviour and cognition. Nature Reviews Neuroscience 10 (6), pp.434–445.

Løken, K.V. 2010. Family income and children's education: using the Norwegian oil boom as a natural experiment. *Labour Economics*, 17 (1), pp.118–129.

Løken, K.V., Mogstad, M., & Wiswall, M. 2012. What linear estimators miss: the effects of family income on child outcomes. *American Economic Journal: Applied Economics*, 4 (2), pp.1–35.

Løken, K.V., Lundberg, S., & Riise, J. 2017. Lifting the burden: formal care of the elderly and labor supply of adult children. *Journal of Human Resources*, 52, pp.247–271.

Pettersen, S.V. 2003. Barnefamiliers tilsynsordninger, yrkesdeltagelse og bruk av kontantstøtte, våren 2002. Rapporter 2003/9, Statistics Norway.

Rivkin, S.G., Hanushek, E.A, & Kain, J.F. 2005. Teachers, schools, and academic achievement. *Econometrica*, 73 (2), pp.417--458.

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

Shonkoff, J.P. & Phillips, D.A (eds). 2000. *From neurons to neighborhoods: the science of early childhood development*. Washington, DC: National Academies Press.

Staiger, D.O. & Rockoff, J.E. 2010. Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives*, 24 (3), pp.97–117.

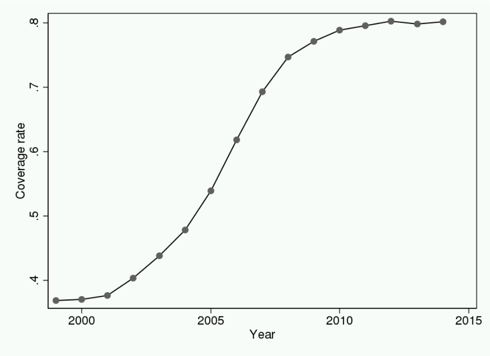
Vassenden, A., Thygesen, J., Bayer, S.B., Alvestad, M., & Abrahamsen, G. 2011. Barnehagens organisering og strukturelle faktorers betydning for kvalitet", Rapport IRIS 2011/029. [Online]. https://www.udir.no/globalassets/upload/barnehage/forskning_og_statistikk/rapporter/strukturelle_faktorers_betydning_for_kvalitet.pdf. Accessed 30th May 2017.

Walters, C.R. 2015. Inputs in the production of early childhood human capital: evidence from Head Start. *American Economic Journal: Applied Economics*, 7 (4), pp.76–102.

Wong, V.C., Cook, T.D., Barnett, W.S., & Jung, K. 2008. An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, 27 (1), pp.122–154.

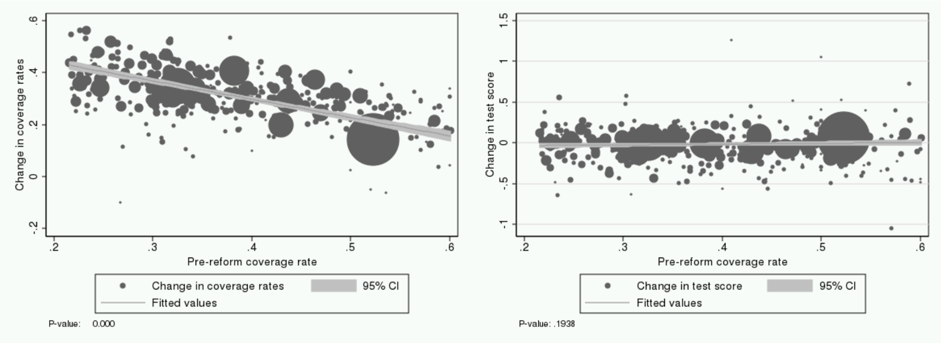
Figures

Figure 1 Coverage by year



Notes: Figure show childcare coverage measured at the end of year in official municipal statistics. A description of the data are found in Section

Figure 2 Reform effect



Notes: Figure a show change in coverage rate for 1-2 year old’s from 2002-2007 by pre reform coverage rate (measured for 1-2 year old’s in 2001). Figure b show change in test score from pre reform cohorts born 1998-2000 to post reform cohorts born 2003-2004 by pre reform coverage rates (Long ITT). Size of dots indicate cohort size of children in municipalities.

Tables

Table 1 - What predicts the childcare expansion?

Dependent variable:	Childcare coverage increase (2002-2007)
	(1)
Pre reform municipal level:	
Childcare coverage rate	-0.475***
	(0.049)
Female employment	-0.114
	(0.194)
Male employment	-0.111
	(0.255)
Size	0.000
	(0.000)
Constant	0.661***
	(0.220)
r2	0.243
N	420

Notes: Table show results from a regression of childcare coverage increase for 1-2 year old's on pre reform coverage and other municipal level characteristics. Pre reform coverage of 1-2 year old's measured in 2001 are used. Municipal level employment by gender are derived from registry data. Employment for persons between 25-39 in each municipality is measured. A person is employed if registered as working more than 100 hours that year. Size is the number of children born in the municipality in 1999. * p<0.10, ** p<0.05, *** p<0.01

Table 2 - Effect of expansion on school performance

Dep var	Average test score		Norwegian	English	Math
	(1)	(2)	(3)	(4)	(5)
Short ITT	-0.095	-0.099	-0.049	0.089	-0.121**
	(0.062)	(0.063)	(0.054)	(0.128)	(0.056)
Long ITT	0.104	0.083	0.186	0.072	-0.039
	(0.149)	(0.129)	(0.121)	(0.098)	(0.108)
Pre reform mean	0.011	0.011	0.026	0.037	0.029
Pre reform sd	0.992	0.992	0.956	0.940	0.962
R squared	0.034	0.117	0.108	0.069	0.113
N	318473	318473	315814	270947	315587
Indiv. controls		x	x	x	x
Munic. Dummies	x	x	x	x	x
Cohort Dummies	x	x	x	x	x

Notes: Table show ITT estimates from Equation 1. Dependent variable is average national standardized national test score at age 10 in Column 1-2, and subject specific test score in Column 3-5. Control variables are gender, mother age, father age, immigrant status, parents labor participation pre birth, parental years of education pre birth. Cohort- and municipality dummies included in all specifications. Standard errors clustered at municipal level in Column 1-3. * p<0.10, ** p<0.05, *** p<0.01

Table 3 - Robustness

	Baseline (1)	Excl. cities (2)	Excl. small (3)	Drop-15/85+ (4)	Excl. 1998 (5)	Excl. 2003 (6)	Flexible trends (7)
Short ITT	-0.099	-0.052	-0.098	-0.152**	-0.119	-0.096	-0.017
	(0.063)	(0.082)	(0.065)	(0.064)	(0.079)	(0.063)	(0.084)
Long ITT	0.083	-0.125	0.119	0.082	0.061	0.098	-0.037
	(0.129)	(0.094)	(0.126)	(0.154)	(0.111)	(0.161)	(0.105)
R-squared	0.116	0.092	0.115	0.117	0.116	0.115	0.118
N	318473	221111	298959	300099	272544	272177	318473
Ind. contr.	x	x	x	x	x	x	x
Munic. dum.	x	x	x	x	x	x	x
Cohort dum.	x	x	x	x	x	x	x

Notes: Table show ITT estimates from Equation 1. Dependent variable is average test score. Column 1 show baseline, Column 2 excludes the 6 largest municipalities, while Column 3 exclude the smallest(less than 300 obs). Column 4 excludes municipalities with pre-reform coverage rates below or above 15 or 85 percentile in the pre coverage rate distribution. Column 5 and 6 exclude cohort 1999 and 2003 respectively. Column 7 interacts predetermined municipal characteristics with cohort dummies to check if municipalities with different observable characteristics have different trends. Control variables are gender, mother age, father age, immigrant status, parents labor participation pre birth, parental years of education pre birth. Cohort- and municipality dummies included in all specifications. Standard errors clustered at municipal level. * p<0.10, ** p<0.05, *** p<0.01

Table 4 - Childcare quality

		Group size		Education level of staff		Costs	
	Proportion children aged 1-2	Children /staff	Adj. care hours /staff	Proportion preschool teachers	Proportion preschool teachers of pedag. leaders	Cost /child	Cost /adj. care hour
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Short ITT	-0.108***	0.799***	-220.386	0.005	-0.004	-6576.222	0.354
	(0.016)	(0.158)	(264.561)	(0.008)	(0.016)	(4223.175)	(2.219)
Long ITT	-0.221***	1.656***	-225.893	-0.003	-0.054	-38100.452***	-6.058**
	(0.026)	(0.319)	(391.395)	(0.024)	(0.044)	(8904.850)	(3.008)
Pre reform mean	0.34	4.66	10951.30	0.33	0.90	92099.60	37.28
Pre reform sd	0.04	0.57	868.04	0.05	0.08	14863.72	5.31
Short ITT/mean	-0.31	0.17	-0.02	0.02	-0.00	-0.07	0.01
Long ITT/mean	-0.64	0.36	-0.02	-0.01	-0.06	-0.41	-0.16
Short ITT/sd	-2.78	1.41	-0.25	0.11	-0.05	-0.44	0.07
Long ITT/sd	-5.68	2.92	-0.26	-0.07	-0.69	-2.56	-1.14
N	275448	275495	266019	274900	275423	265715	265715

Notes: Table show estimate of Equation 2 using different measures of municipality level child care quality as outcomes. Dependent variable is average municipal level childcare quality from the year children in each cohort turn 2, until the year it turns 5. “Short” is a indicator for cohorts born 2001-2002 that are only partly influenced by the reform, while “Long” are indicator for cohorts born 2003-2004 that are more heavily influenced by the reform. “Short ITT” are the “Short ” indicator interacted with pre reform municipal level home care coverage rates, while “Long ITT” are “Long” indicator interacted with pre reform coverage rates. Means and standard deviation of quality measures are given, as well as the ITT-estimates size relative to these. Childcare quality information is missing from some municipalities in some years. Standard errors are clustered at municipality level. * p<0.10, ** p<0.05, *** p<0.01

Table 5 - Effect of expansion by municipality type

Dep. var.	Proportion preschool teachers			Proportion preschool teachers of pedag. leaders		
	Low	Mid	High	Low	Mid	High
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A) - Dep. var. - Average test score						
Short ITT	-0.041	0.063	-0.317**	-0.039	0.012	-0.112
	(0.073)	(0.113)	(0.154)	(0.083)	(0.113)	(0.168)
Long ITT	0.281*	-0.009	-0.347*	0.283*	0.057	-0.409**
	(0.156)	(0.145)	(0.189)	(0.163)	(0.139)	(0.165)
N	116539	104198	97605	107208	116806	94459
Panel B) - Dep. var. - Norwegian						
Short ITT	0.041	0.069	-0.285*	0.009	0.044	-0.081
	(0.062)	(0.101)	(0.151)	(0.079)	(0.103)	(0.159)
Long ITT	0.380***	0.006	-0.179	0.354***	0.047	-0.155
	(0.115)	(0.112)	(0.160)	(0.121)	(0.115)	(0.132)
Panel C) - Dep. var. - English						
Short ITT	0.314**	0.009	-0.380*	0.306*	0.062	-0.212
	(0.140)	(0.143)	(0.215)	(0.178)	(0.152)	(0.207)
Long ITT	0.225	0.085	-0.248	0.228	0.139	-0.284*
	(0.143)	(0.148)	(0.168)	(0.152)	(0.134)	(0.156)
Panel D) - Dep. var. - Math						
Short ITT	-0.073	-0.037	-0.252*	-0.037	-0.091	-0.147
	(0.082)	(0.101)	(0.127)	(0.092)	(0.098)	(0.130)
Long ITT	0.101	-0.072	-0.402*	0.133	-0.028	-0.504***
	(0.139)	(0.136)	(0.205)	(0.142)	(0.142)	(0.172)

Notes: Table show ITT estimates from Equation 1 split into 4 panels. Dependent variable are average test score, Norwegian score, English score and maths core in panel a), b), c) and d) respectively. In Column 1-3 Equation 1 is estimated separately by dividing municipalities according to pre reform proportion preschool teachers municipal distribution quantile 1-3. In Column 4-6 the same is done with proportion preschool teachers of pedagogical leaders, while Column 7-9 show the results for proportion pedagogical leaders. Individual controls listed in section 3, cohort dummies and municipality dummies included in all specifications. Standard errors clustered at municipal level. * p<0.10, ** p<0.05, *** p<0.01

Table 6 - Mechanism

	Municipality level quality			
	Low	Mid	High	N
	(1)	(2)	(3)	(4)
Panel A)				
Children /staff	2.388***	1.252***	0.788*	275495
	(0.343)	(0.336)	(0.403)	
Adj. care hours /staff	167.904	414.792	-960.935	266019
	(653.280)	(868.518)	(780.747)	
Prop.pre teachers	-0.033	0.017	0.057	274900
	(0.026)	(0.039)	(0.035)	
Prop. pre teachers	-0.117**	0.015	-0.029	275423
of pedag. leaders	(0.056)	(0.079)	(0.057)	
Cost /child	-38168.154***	-39231.407**	-15010.023	265715
	(9995.699)	(17703.621)	(11363.344)	
Cost /adj. care hour	-2.462	-12.169*	-1.195	265715
	(5.935)	(6.822)	(3.978)	
Panel B)				
Mother LFP	-0.092**	-0.052	-0.084**	321684
	(0.037)	(0.045)	(0.040)	
Father LFP	-0.039	0.008	-0.013	321684
	(0.029)	(0.019)	(0.042)	
Panel C)				
Months before age 3	-8.609***	-3.587**	-2.536	321684
	(2.003)	(1.576)	(1.806)	
Hours before age 3	-906.587***	-336.271**	-199.824	321684
	(224.604)	(158.986)	(189.619)	
Coverage rate 1-2	-0.665***	-0.239***	-0.140*	275495
	(0.048)	(0.060)	(0.079)	
Coverage rate 3-5	-0.192***	-0.258***	-0.358***	275495
	(0.031)	(0.038)	(0.062)	
Panel D)				
Average test score	0.283*	0.057	-0.409**	321684
	(0.163)	(0.139)	(0.165)	

Notes: Table show long run ITT estimates from Equation 1 split into 4 panels. Dependent variable are indicated in row header. Reform effect on measures of childcare quality, parental labor force participation, childcare attendance and test score are shown in panel a), b), c) and d) respectively. Equation 1 is estimated separately by dividing municipalities according to quantiles in the pre reform proportion preschool teachers municipal distribution. Total sample size indicated in the column most to the right. Control variables are are included in individual level regressions and listed in Section 3. Cohort dummies and municipality dummies included in all specifications. Standard errors clustered at municipal level. * p<0.10, ** p<0.05, *** p<0.01

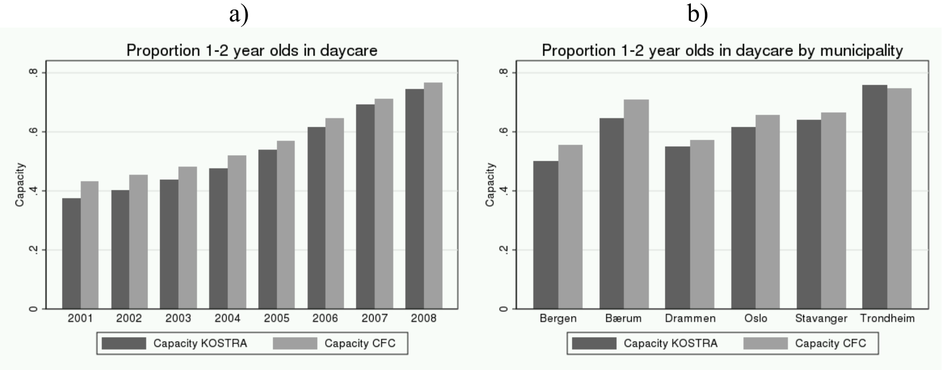
Table 7 - Heterogeneity by child characteristics

	Female	Male	Low-SES	High-SES
	(1)	(2)	(3)	(4)
Panel A) - Low quality municipalities				
Short ITT	0.030	-0.104	-0.058	-0.043
	(0.092)	(0.109)	(0.149)	(0.096)
Long ITT	0.325*	0.229	0.646***	0.177
	(0.183)	(0.178)	(0.130)	(0.196)
R-squared	0.147	0.155	0.145	0.145
N	53815	53393	24715	82493
Panel B) - Medium quality municipalities				
Short ITT	-0.029	0.064	-0.114	0.050
	(0.139)	(0.138)	(0.159)	(0.123)
Long ITT	0.067	0.051	0.188	0.030
	(0.147)	(0.158)	(0.165)	(0.155)
R-squared	0.098	0.095	0.100	0.092
N	58244	58562	26131	90675
Panel C) - High quality municipalities				
Short ITT	-0.072	-0.133	-0.107	-0.130
	(0.181)	(0.250)	(0.201)	(0.178)
Long ITT	-0.463***	-0.356	0.050	-0.577***
	(0.153)	(0.249)	(0.139)	(0.150)
R-squared	0.093	0.094	0.102	0.088
N	47052	47407	23552	70907
Ind. contr.	x	x	x	x
Munic. dum.	x	x	x	x
Cohort dum.	x	x	x	x

Notes: Table show ITT estimates from Equation 1. Dependent variable is average test score. Panel A, B and C show estimates on low, medium and high quality municipalities as measured by proportion preschool teachers of pedagogical leaders. Control variables are gender, mother age, father age, immigrant status, parents labor participation pre birth, parental years of education pre birth. Cohort- and municipality dummies included in all specifications. Standard errors clustered at municipal level. * p<0.10, ** p<0.05, *** p<0.01

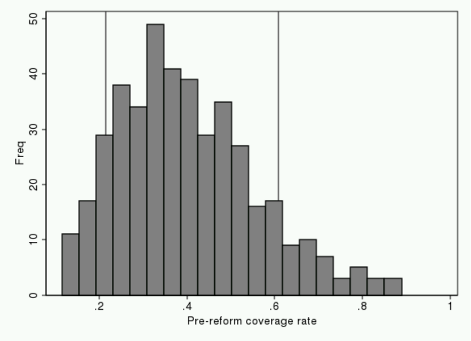
Appendix Figures

Figure A1 KOSTRA and CFC coverage rates



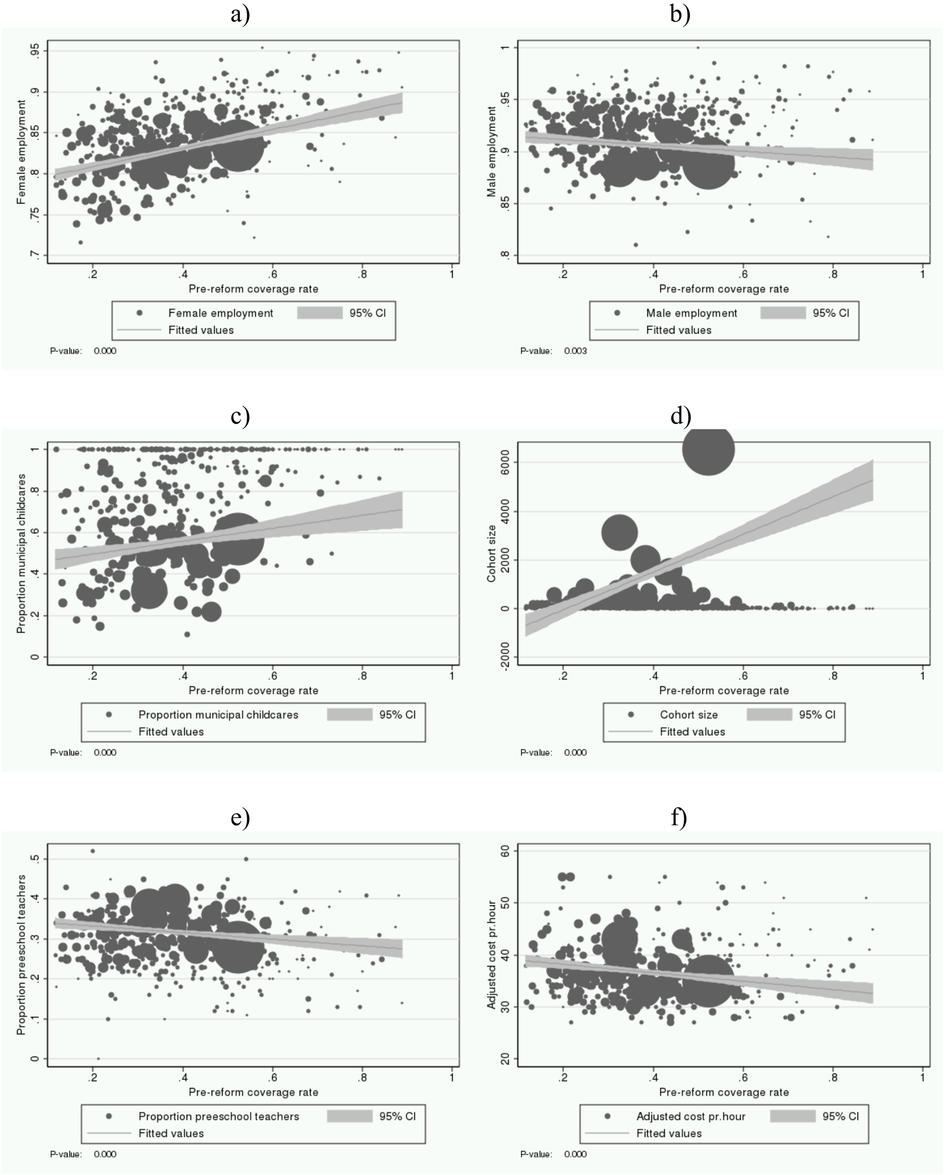
Notes: a) compares KOSTRA numbers to CFC numbers across years. CFC numbers are based on counting the number of children in childcare using the CFC database. CFC numbers are at the individual level. KOSTRA numbers are based on childcare centers' reporting the number of one- and two-year-olds in care to municipalities, which the municipalities then report to Statistics Norway; b) compares KOSTRA numbers with CFC numbers across the six largest municipalities in Norway for 2006. Correlations between CFC and KOSTRA across municipalities by year from 2000 through 2008 are as follows: 0.94, 0.94, 0.93, 0.94, 0.93, 0.89, 0.86, 0.87, 0.83.

Figure A2 – Pre-reform coverage distribution



Note: The lines indicate where the sample has been cut to exclude municipalities below the 10th and above the 90th percentiles; coverage rate measured in 2001.

Figure A3 Municipality descriptors



Notes: Figures show descriptions of municipalities; P-values show the significance of the slopes. Numbers are measured in 2001.

Appendix Tables

Table A1 - Data and insitutional details

	KOSTRA rates	CFC data	Exams 5th grade*	Exams 8th grade	Attendance 1-2	Reform
	(1)	(2)	(3)	(4)	(5)	(6)
Cohort						
1998		x	x	x	1999-2001	Pre
1999	x	x	x	x	2000-2002	Pre
2000	x	x	x	x	2001-2003	Pre
2001	x	x	x	x	2002-2004	Phase-in
2002	x	x	x		2003-2005	Phase-in
2003	x	x	x		2004-2006	Post
2004	x	x	x		2005-2007	Post
2005	x	x			2006-2008	
2006	x	x			2007-2009	
2007	x	x			2008-2010	
2008	x				2009-2011	
2009	x				2010-2012	
2010	x				2011-2013	

Notes: Table describe data and institutional details. *Math and English tests are electronically corrected from cohort 2000 and onward. *Lack English test scores for cohort born 2001 due to test being canceled.

Table A2 - Robusness II

	Baseline	Excl. cities	Excl. small	Drop-15/85+	Excl. 1998	Excl. 2003	No controls	lexible trends
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A) - Low quality municipalities								
Short ITT	-0.039	-0.064	-0.044	-0.037	-0.055	-0.037	-0.009	-0.128
	(0.083)	(0.137)	(0.088)	(0.091)	(0.093)	(0.083)	(0.081)	(0.141)
Long ITT	0.283*	-0.026	0.315*	0.314*	0.262*	0.411**	0.344*	-0.063
	(0.163)	(0.192)	(0.161)	(0.187)	(0.158)	(0.194)	(0.182)	(0.203)
R-squared	0.149	0.100	0.148	0.150	0.150	0.148	0.052	0.151
N	107208	56082	102299	101506	91942	91279	107208	107208
Panel B) - Medium quality municipalities								
Short ITT	0.012	0.034	0.005	-0.134	0.094	0.016	0.010	0.071
	(0.113)	(0.117)	(0.114)	(0.131)	(0.121)	(0.113)	(0.119)	(0.118)
Long ITT	0.057	-0.019	0.055	0.032	0.140	0.014	0.086	0.130
	(0.139)	(0.129)	(0.141)	(0.169)	(0.150)	(0.185)	(0.150)	(0.150)
R-squared	0.095	0.091	0.094	0.094	0.095	0.093	0.019	0.096
N	116806	102943	115704	107158	99768	99860	116806	116806
Panel C) - High quality municipalities								
Short ITT	-0.112	-0.182	-0.190	-0.129	-0.171	-0.109	-0.158	-0.301
	(0.168)	(0.168)	(0.170)	(0.186)	(0.165)	(0.167)	(0.169)	(0.195)
Long ITT	-0.409**	-0.425**	-0.400**	-0.469***	-0.466***	-0.582***	-0.468**	-0.400**
	(0.165)	(0.172)	(0.183)	(0.175)	(0.157)	(0.211)	(0.180)	(0.197)
R-squared	0.092	0.088	0.090	0.091	0.091	0.093	0.016	0.094
N	94459	62086	91438	91435	80834	81038	94459	94459
Ind. contr.	x	x	x	x	x	x		x
Munic. dum.	x	x	x	x	x	x	x	x
Cohort dum.	x	x	x	x	x	x	x	x

Notes: Table show ITT estimates from Equation 1. Dependent variable is average test score. Panel A, B and C show estimates on low, medium and high quality municipalities as measured by proportion preschool teachers of pedagogical leaders. Column 1 show baseline, Column 2 excludes the 6 largest municipalities, while Column 3 exclude the smallest(less than 300 obs). Column 4 excludes municipalities with pre reform coverage rates below or above 15 or 85 percentile in the pre coverage rate distribution. Column 5 and 6 exclude cohort 1999 and 2003 respectively. Column 7 interacts predetermined municipal characteristics with cohort dummies to check if municipalities with different observable characteristics have different trends. Control variables are gender, mother age, father age, immigrant status, parents labor participation pre birth, parental years of education pre birth. Cohort- and municipality dummies included in all specifications. Standard errors clustered at municipal level. * p<0.10, ** p<0.05, *** p<0.01

Table A3 - Testing for different change in high- and low-quality municipalities

	Children /staff (1)	Adj. care hours /staff (2)	Proportion preschool teachers (3)	Proportion preschool teachers of pedag. leade (4)	Cost /child teachers (5)	Cost /adj care hour pr. child (6)	Maternal LFP (7)
Panel A							
Diff	-1.601***	-1193.466	-0.030	0.086	23375.934	1.476	0.006
	(0.528)	(1039.278)	(0.046)	(0.079)	(15248.707)	(7.205)	(0.054)
N	174711	165964	186179	174639	165660	165660	203758

Panel B	Paternal LFP (8)	Months (9)	Hours (10)	Coverage rate 1-2 (11)	Coverage rate 3-5 (12)	Average test score (13)
Diff	0.027	5.922**	690.523**	0.524***	-0.155*	-0.695***
	(0.051)	(2.656)	(289.534)	(0.157)	(0.088)	(0.237)
N	203758	203758	203758	174711	174711	201667

Notes: Table show estimates of γ(9) from estimating:

$$\begin{aligned}&Outcome_{it} = \gamma_1 + \gamma_2 Short_{it} + \gamma_3 Long_{it} + \gamma_4 (Short_{it} \cdot Highquality_{it}) + \gamma_5 (Long_{it} \cdot Highquality_{it}) +\\ &\gamma_6 (Short_{it} \cdot \Pr eCoverage_{it}) + \gamma_7 (Long_{it} \cdot \Pr eCoverage_{it}) + \gamma_8 (Short_{it} \cdot \Pr eCoverage_{\mathbf{it}} \cdot Highquality_{it})\\ &+\gamma_9 (Long_{it} \cdot \Pr eCoverage_{\mathbf{it}} \cdot Highquality_{it}) + \gamma_{10} + \gamma_8\end{aligned}$$

Highquality(it) is a dummy indicating 1 if the child was born in a high-quality municipality, and 0 if it belongs to a low quality municipality. The sample consists of only high and low quality municipalities. * p<0.10, ** p<0.05, *** p<0.01

Table A4 - Parental labor force participation

	Mother		Father	
	(1)	(2)	(3)	(4)
Short ITT	0.008	-0.004	-0.011	-0.024
	(0.018)	(0.024)	(0.011)	(0.015)
Long ITT	-0.055***	-0.082***	-0.012	-0.028
	(0.021)	(0.030)	(0.015)	(0.025)
Dep. var mean	0.537	0.537	0.879	0.879
Dep. var sd	0.499	0.499	0.326	0.326
R2	0.030	0.309	0.010	0.283
N	321684	321684	321684	321684
Indiv. controls		x		x
Munic. Dummies	x	x	x	x
Cohort Dummies	x	x	x	x

Notes: Table show ITT estimates from Equation 1. Dependent variable is parental labor force participation. Pre coverage are the childcare coverage rate for 1-2 year old's in the child's birth municipality registered in year 2001. Control variables are gender, mother age, father age, immigrant status, parents labor participation pre birth, parental years of education pre birth. Cohort- and municipality dummies included in all specifications. Standard errors clustered at municipal level in Column 1-3. * p<0.10, ** p<0.05, *** p<0.01

Table A5 - Mechanism II

Pre reform indicator	Coverage rate 1-5 year old's			N
	Low	Mid	High	
	(1)	(2)	(3)	(4)
Panel a)				
Child/staff	2.089***	0.226	0.017	275495
	(0.767)	(0.392)	(0.484)	
Adj. care hour/staff	-219.260	-563.015	-1358.814	266019
	(1016.264)	(573.476)	(1054.372)	
Prop. pre teachers	-0.031	0.058	-0.019	275423
of pedag. leaders	(0.094)	(0.092)	(0.065)	
Prop.pre teachers	-0.001	0.039	0.048	274900
	(0.039)	(0.043)	(0.040)	
Costs pr child	-34152.638**	-18409.958	-4829.755	265715
	(14399.130)	(15027.284)	(13464.052)	
Costs pr adj. care hour	-0.717	-4.706	-0.369	265715
	(6.640)	(5.640)	(5.299)	
Panel b)				
Mother LFP	-0.131**	-0.087*	-0.078*	321684
	(0.058)	(0.049)	(0.042)	
Father LFP	-0.032	-0.001	0.015	321684
	(0.041)	(0.016)	(0.039)	
Panel c)				
Months	-9.639***	-3.827**	-1.125	321684
	(3.413)	(1.769)	(1.984)	
Hours	-997.015***	-374.792**	-53.488	321684
	(375.889)	(177.379)	(206.763)	
Coverage rate 1-2	-0.822***	-0.239***	-0.015	275495
	(0.071)	(0.065)	(0.082)	
Coverage rate 3-5	-0.409***	-0.467***	-0.505***	275495
	(0.043)	(0.038)	(0.062)	
Panel d)				
Average test score	0.116	0.117	-0.478***	321684
	(0.275)	(0.153)	(0.149)	

Notes: Table show long run ITT estimates from Equation 1 split into 4 panels. Dependent variable are indicated in row header. Reform effect on measures of childcare quality, parental labor force participation, childcare attendance and test score are shown in panel a), b), c) and d) respectively. Equation 1 is estimated separately by dividing municipalities according to quantiles in the pre reform proportion preschool teachers municipal distribution. Pre reform coverage rates for 1-5 year old's are used as pre reform indicator of intensity of expansion. Total sample size indicated in the column most to the right. Control variables are are included in individual level regressions and listed in Section 3. Cohort dummies and municipality dummies included in all specifications. Standard errors clustered at municipal level. * p<0.10, ** p<0.05, *** p<0.01

Table A6 - Difference between municipality types

	Prop. 1-2 y old's (1)	Pre coverage 1-2 y. old's (2)	Pre coverage 3-5 y. old's (3)	Prop. pre teachers/ pedag. (4)	Prop. pre teachers (5)	Cost pr. child (6)	Adj. cost pr. care hour (7)
Panel A							
High	-0.050*	-0.076***	-0.047***	0.210***	0.091***	961.963	1.362*
	(0.029)	(0.012)	(0.010)	(0.007)	(0.005)	(1554.302)	(0.805)
Dep. var. mean	0.486	0.386	0.798	0.886	0.319	79531.650	34.446
Dep. var. sd	0.194	0.096	0.088	0.098	0.050	13098.657	5.676
R2	0.013	0.149	0.081	0.793	0.624	0.002	0.013
N	231	231	231	231	229	226	226

Panel B	Children /staff (9)	Children pr. pedag. leader (10)	Adj. care hours/ staff (11)	Prop. munic. childcare (11)
High	0.301***	0.115	-60.110	-0.024
	(0.087)	(0.521)	(118.559)	(0.028)
Dep. var. mean	4.773	18.090	10846.398	0.524
Dep. var. sd	0.660	3.706	928.824	0.202
R2	0.050	0.000	0.001	0.003
N	231	230	227	229

Notes: Table show results from regressions of municipal characteristics on indicator for municipality with a high proportion of preschool teachers of pedagogical leaders. Only municipalities with highest and lowest proportion are included in the sample.
* p<0.10, ** p<0.05, *** p<0.01

Table A7 - Robustness III - Low SES children, low-quality municipalities

	Baseline	Excl. cities	Excl. small	Drop-15/85+	Excl. 1998	Excl. 2003	No controls	Flexible trends
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Short ITT	-0.056	0.134	-0.080	-0.148	-0.048	-0.071	-0.085	0.130
	(0.148)	(0.259)	(0.144)	(0.136)	(0.177)	(0.148)	(0.153)	(0.273)
Long ITT	0.651***	0.692***	0.681***	0.562***	0.660***	0.803***	0.630***	0.760***
	(0.130)	(0.252)	(0.131)	(0.142)	(0.145)	(0.166)	(0.124)	(0.271)
R-squared	0.146	0.107	0.146	0.148	0.146	0.148	0.030	0.148
N	24715	12235	23132	23506	20877	21177	24715	24715
Ind. contr.	x	x	x	x	x	x		x
Munic. dum.	x	x	x	x	x	x	x	x
Cohort dum.	x	x	x	x	x	x	x	x

Notes: Table show ITT estimates from Equation 1. Dependent variable is average test score, and sample is restricted to low-SES children. The table show estimates on low quality municipalities as measured by proportion preschool teachers of pedagogical leaders. Column 1 show baseline, Column 2 excludes the 6 largest municipalities, while Column 3 exclude the smallest(less than 300 obs). Column 4 excludes municipalities with pre reform coverage rates below or above 15 or 85 percentile in the pre coverage rate distribution. Column 5 and 6 exclude cohort 1999 and 2003 respectively, while Column 7 perform the estimation without controls. Column 8 interacts predetermined municipal characteristics with cohort dummies to check if municipalities with different observable characteristics have different trends. Control variables are gender, mother age, father age, immigrant status, parents labor participation pre birth, parental years of education pre birth. Cohort- and municipality dummies included in all specifications. Standard errors clustered at municipal level. * p<0.10, ** p<0.05, *** p<0.01

Households' responses to price changes of formal childcare[*]

Leroy Andersland[†] Øivind A. Nilsen[‡]

This version: 13 August, 2017

Abstract

This study examines the changes in childcare attendance following a Norwegian reform that introduced a money transfer to families who did not send their child to formal childcare. This cash-for-care reform raised the price of formal childcare relative to its alternatives by about 108% for 1–2 year olds but not 3–5 year olds. Using household surveys conducted before and after the reform, the analysis reveals that childcare attendance fell by 14.4 percentage points four years after the reform took place. In contrast to previous studies, the results indicate that the most important alternative mode of care to formal childcare is parental care. Furthermore, the main alternative for households of low socioeconomic status is parental/relative care, whereas for high socioeconomic status families the alternatives include day parks and nannies as well as parental care.

Keywords: Public Policy, Cash Incentives, Childcare, Difference-in-Difference

JEL codes: D10, J13, H31

[*] Acknowledgment: This paper has benefitted from comments and suggestions from Ian Fried, Astrid Kunze, Katrine V. Løken, Trond Pedersen, Katrine H. Reiso, Kjell Vaage, Thomas von Brasch, two anonymous referees, from participants in a workshop held at the Norwegian School of Economics, attendees at the 30th European Society for Population Economics Conference in Berlin, and participants at the 39th Annual Meeting of the Norwegian Association of Economists. The usual disclaimer applies.

[†] Department of Economics, University of Bergen, 5020 Bergen, Norway, email; leroy.andersland@econ.uib.no, tel; +47 47313906

[‡] Department of Economics, Norwegian School of Economics, Hellev. 30, 5035 Bergen, Norway, CESifo-Munchen and IZA-Bonn, email; oivind.nilsen@nhh.no

1. Introduction

Nowadays, much childcare takes place outside the family home in many European countries. For the youngest children, attendance rates at formal childcare are especially high in the Nordic countries (OECD, 2016). Extensive public subsidization of the childcare sector has facilitated the expansion in childcare attendance rates, with public funding supported by arguments that formal childcare has beneficial effects for both parental employment and the children themselves.[1]

Parents making decisions on the form of childcare used face a number of considerations. Labor market attachment, childcare quality assessments, childcare availability, and price are all components that can influence the decision process about the form of childcare. Nevertheless, given that formal childcare has beneficial effects for the children and their mothers, it is of great importance for both policy makers and researchers to know how parents respond to price changes in formal childcare. Childcare subsidies are costly to the taxpayers. Furthermore, subgroups in the population might respond differently to price changes. Only a few studies have isolated large shocks to childcare prices unaccompanied by other (non-price) changes and examined their consequences. Evidence on the subject remains scarce because of either a lack of data or suitable natural experiments, and more work on this question is needed. In light of recent emphasis of the importance of childcare and other early influences on later life outcomes, our main contribution is to add evidence on how price sensitive parents are to changes in childcare prices, and show results on the consequences of a price change in formal childcare for other care arrangements.

A possible way to study responses to changes in childcare prices is to analyze changes in childcare subsidies. However, this method has some limitations. For example, childcare

[1] For a discussion on the effects of public childcare on parental employment, see Lundin *et al*. (2008), Mogstad and Havnes (2011), Baker *et al*. (2008), Lefebre and Merrigan (2008), Bettendorf *et al*. (2015) and Bauernschuster and Schlotter (2015). A recent survey on the literature of the effects of childcare on children can be found in Heckman and Mosso (2014).

subsidies tend to be means tested, which complicates the interpretation of any effects, and using subsidy eligibility cut-offs necessarily limits the validity of any effect to some specific subpopulation. As an alternative, we propose to use the introduction of a particular type of policy, namely, a cash-for care (CFC) reform, to examine how households respond to price changes for childcare in a way that does not suffer from the same limitations.[2] The CFC is a pecuniary transfer to parents that do not send their children to formal childcare. The Norwegian data provide promising context to investigate the question of how households respond to price changes in formal childcare. The introduction of the CFC reform was not followed by tax cuts nor transfers, i.e. no simultaneous changes in income. Neither was the reform followed by changes in means testing. Furthermore, the CFC reform did not directly include any changes in capacity. Finally, the introduction of the CFC reform was introduced at the same time in whole Norway. Thus, our analysis does not suffer from potential biases from unobserved factors that may vary if one would do an analysis between different states. The CFC benefit was available in Norway for 1-year-old children from 1 August 1998 and for 2-year-old children from 1 January 1999. The reform would eventually provide 3,000 Norwegian kroner (NOK) per month to parents choosing not to send their 1–2-year-old children to formal childcare providers receiving public funds.[3] As the benefit was unavailable for 3–5 year olds, we employ a difference-in-differences (DID) strategy by comparing the rates of childcare for eligible and ineligible children, before and after the introduction of the reform.

At the time of the reform, parents paid on average 2,775 NOK a month for care in formal childcare; thus, the reform represented a nearly 108% price increase for formal

[2] The notions "cash-for care" and "home care allowance" reforms are used interchangeably in this reform. We will use the notion cash-for-care consistently in this paper.

[3] 1 NOK ≈ 0.125 USD in 2002.

childcare relative to any other forms of childcare, which is quite significant.[4] For this reason, the CFC reform may yield valuable information on how price sensitive parents are to the price of formal childcare in general, as well as what the alternative modes of care is. It may also highlight the price sensitivity of particular subgroups in the population as there may be many reasons why some demographic groups are more likely to receive the CFC benefit and less likely to send their children to formal childcare. This paper attempts to address this issue by separating the effect of the benefit for groups of different socioeconomic status (SES).

Some literature has already estimated the effect of childcare prices on childcare attendance. The most recent contributions use policy reforms or rules that provide exogenous shocks to childcare prices and study its impact.[5] Baker *et al*. (2008) examined a reform in the childcare sector in Quebec that included a generous childcare subsidy that set the price of childcare at just 5 Canadian dollars per day. After comparing childcare attendance in Quebec to the rest of Canada, before and after the reform, Baker *et al*. (2008) found that childcare use increased, while there was a shift from care by relatives and non-licensed non-relatives. No effect on the care in own home was found. An important factor that separates that study from the current analysis is that the childcare subsidy coincided with an expansion in childcare capacity. Therefore, Baker *et al*. (2008) did not isolate the effect of the subsidy on childcare attendance. Another separating feature is that prior to the reform identified in Baker *et al*. (2008), other childcare subsidies depended on family income. The effective price change in childcare following the new program therefore also depended on family income. In contrast, the CFC subsidy in Norway is uniform for all families, which makes it easier to more directly interpret and compare any price responses.

[4] Reppen and Rønning (1999) report the average monthly payment for formal childcare when a 1-2 year old child is cared for outside the home.

[5] An earlier literature estimates the price elasticities of childcare, including Blau and Hagy (1998), Powell (2002) and Connelly and Kimmel (2003). These studies report price elasticities ranging from –0.3 to –1.0. A contribution of this analysis relative to that literature is the use of a different identification strategy.

Gathmann and Sass (2017) is closest in spirit to the present analysis because it also used a nationwide population survey to analyze the consequences of the introduction of a CFC program, but in a single German state. As the benefit applied in only one state, a factor that separates our studies is that Gathmann and Sass (2017) mainly compared the childcare outcomes in the reform state relative to those in other states, whereas we compared the childcare outcomes for eligible and ineligible children across different ages. In addition, our survey data contains information of individual childcare prices, which allows us to perform a detailed analysis of price responses of different groups.

As an alternative, Black *et al*. (2014) considered the consequences of childcare subsidies by utilizing the fact that eligibility depends on sharp family income cut-offs. By comparing families immediately below and right above the income cut-offs, they found among other things, no effect of the subsidy on formal childcare attendance for children aged 5 years. One explanation for this finding is that there is an excess demand (or rationing) for childcare. It is then not the price that is important, but the availability of a spot. Another possible explanation is that information about the subsidy is not easily available to parents before they actually apply for childcare. Lastly, an important point when comparing the analysis in Black *et al*. (2014) to ours is that the subsidy eligibility cut-offs they considered were for 5-year-old children, while the children in our study are much younger (1–2 year-olds).

Other studies that have specifically looked at the Norwegian CFC reform have mostly focused on the effects of maternal labor force participation.[6] For instance, Schøne (2004) associated a modest reduction in the female labor supply with the reform, while Naz (2004) identified a relatively larger labor participation response among more highly educated mothers. Kornstad and Thoresen (2007) build a simulation model of households' utility-

[6] This is in line with an international literature. Examples are Blau and Robins (1988), Leibowitz *et al*. (1992), Lundholm and Ohlsson (1998), Ribar (1992), and Tekin (2005, 2007).

maximization under budget constraints, and find that mothers reduce their labor supply by about 9% as a result of the CFC reform. Hardoy and Schøne (2008) focus on the labor supply of non-Western immigrant females and find that the CFC reform reduced immigrant female labor supply more than it did for non-immigrant females. This suggests that immigrants are more responsive to the reform on the labor supply margin. Drange and Rege (2013) look at long-term outcomes and find that the effect on mothers labor market outcomes persist even after the children become CFC ineligible, but disappears when the children are aged 6-7. Bettinger *et al*. (2014) explores what happens to older siblings of CFC eligible children, while Drange (2015) focus on both parents time allocation.

Other Norwegian reforms have also been used to investigate labor supply effects of childcare. Hardoy and Schøne (2015) use the so-called ''Childcare Centre Agreement'' (Barnehageforliket) effective from April 2004. This was a broad political consensus agreement reached in 2003 that included reduced costs and increased capacity. The results indicate a smaller sensitivity to prices than other studies. Kalb and Thoresen (2010) use the same reform as a basis for a simulation study. They conclude that the both the female labor supply effect and income redistributional performance of fee reductions is weak, and that appears to be relatively little gain at a rather high cost. Finseraas *et al*. (2017) look at a school starting age reform of 1997 - later on referred to as Reform 97. Their findings indicate strongest effects among mothers with low wage potential.

Of course, the impact of the relative price increase in formal childcare may affect attendance at other childcare alternatives. There are two main motivations for knowing these alternative modes. One reason may be that a relative price increase in childcare can have direct effects on the labor market attendance of mothers. If an important alternative to formal childcare is nanny care, then the employment effects of childcare prices on mothers are not clear. A second reason is that knowing about the alternative modes of care improves the interpretation of the effects of formal childcare attendance on children's future outcomes. Given the discussion in the literature on the effect of early intervention, it is important to know the main alternative(s) for formal childcare. Since this literature often estimates and compares effects by parental socioeconomic status, knowing the alternative for both groups are important for the interpretation.

This study contributes to the literature by assessing how childcare attendance changes as a response to price changes in formal childcare. A particular contribution is the exploration of the alternative modes of care, as they may be different for children of different age groups, or social-economic groups. Alternative modes can also be influenced by the specific country/institutional contexts. In contrast to other studies of the Norwegian cash for care reform; we focus on the effects on the children.

Three main findings arise from the analysis. First, the results demonstrate that the price change reduced usage among eligible children by 14.4 percentage points by 2002. This points to a childcare price elasticity of about –0.25. Second, the price change affected attendance most among the youngest CFC eligible children. Lastly, the main alternative mode of care to formal childcare is parental care. While alternative mode of care for households of low socioeconomic status is parental/relative care, for high socioeconomics status families the alternatives include day parks and nannies as well as parental care.

The remainder of the paper is organized as follows. Section 2 discusses the institutional framework for the CFC policy, while Section 3 describes the data and details the econometric model used in the analysis. Section 4 reports and discusses the results, and Section 5 performs robustness checks. Finally, Section 6 concludes.

2. Institutional Setting

The development of a public childcare sector in Norway relates to the increase in female labor force participation. In the mid-1960s, few mothers were active labor market participants, and correspondingly there were relatively few childcare centers. As female labor participation accelerated in the 1970s, there was a corresponding increase in childcare attendance. From 1975 to 2002 labor force participation of females aged 25-54 increased from 51.2% to 80.7% (Statistics Norway, 2017). In 2002 about 43% held part time positions, while 57% had full time positions. The proportions have stayed about the same across years. In the same time period, the childcare coverage of 1-5 year old children increased from 7% to 66% (NSD 2017).

In 1997 the authorities implementated the so-called Reform 97. Together with changes in the school curriculum, the law meant an expansion of compulsory schooling from nine to ten years, and a requirement as of August 1997 for all children to start school at the age of six (Norwegian Ministry of Education, 1996). A consequence of this law change was that almost no 6 year olds were registered in formal childcare by the end of 1997.

The introduction of the CFC benefit was clearly planned in a new government coalition political platform signed in October 1997 (Christian-Green-Liberal coalition political party platform; Voksenåserklæringen). The parliament passed the law that would implement the CFC benefit in April 1998. There were three main purposes of the CFC reform: (i) provide more freedom of choice to parents of the form of childcare, (ii) provide parents with more time to be with their children, and (iii) to redistribute funds to families that did not receive services from subsidized childcare providers. The CFC benefit was available for 1-year-old children from 1 August 1998 and for 2-year-old children from 1 January 1999. It is paid to parents - with whom the child lives – and who do not send their eligible child to public subsidized childcare. Parents need to apply for the benefit. The benefit was initially set to 3,000 NOK per month from 1 August 1998. From 1 January 1999 it decreased to 2,263 NOK, before it increased back to 3,000 NOK from 1 January 2000 to 1 August 2003 (Bakken and Myklebø, 2010). Families who use childcare part time would be compensated according to the fraction of a full daycare seat used. There have been some changes to the law over time concerning the pecuniary generosity and age criteria, but the main features of the benefit have remained largely the same.

In Norway, formal public or private childcare is centrally regulated through the Kindergarten Act and by different prescripts to the act. This provides a set of common rules for childcare across Norway, as childcare centers are administered at the municipal level. The maximum number of children in full time care per pedagogical leader (with required certified education) is 9 for children below 3 and 18 for children aged 3-6. There is a stated norm of 1 caregiver per 3 children for children below 3 and 1 caregiver per 6 children above 3. In 2002 about 42% of children attended privately owned childcare centers. Privately owned childcare centers still receive public subsidies.

Outside the family, ordinary childcare centers, family childcare, relatives, nannies or day parks normally care for children in Norway. The professional alternatives can have private or public ownership, but all types of childcare receive operating funds from public sources, except for nannies and day parks. Family childcares are usually smaller groups where the care is run by one of the parents in private homes. Nannies are privately operating childminders that are not subject to the same regulations as childcares that receive public funding. Since the start of the integration of mothers into the labor market, there has been a significant use of nannies. The peak is considered to be around 1989 when 22% of all parents reported use of nannies (Blix and Guldbrandsen, 1992). Possible due to the growing public childcare sector, nanny use decreased in the beginning of the 1990s (Blix and Guldbransen, 1993). Because of less oversight, information about the nanny-market is scarce. Based on response in the household surveys, nannies seem to have few and young children in care (Bakelien *et al*. 2001). Day parks are organized as outdoor playgroups. They operate with shorter opening hours than regular childcare centers, and do not receive central government public subsidies. As a consequence, it is possible to receive full CFC benefit and at the same time use day parks

Formal childcare is financed through central government subsidies, municipal subsidies and parental co-payments, hereinafter referred to as price. At the time of the reform, the level of the price was not regulated in the Kindergarten act. Therefore the owners of the childcare generally could set the level themselves.[7] Reppen and Rønning (1999, Table 2.15) show that households that use formal childcare as care alternative outside the home paid on average 2,775 NOK a month for care of 1-2 year olds. Parents that used nannies paid on average 2,707 NOK a month. The report also provides estimates of the hourly expense of care of different childcare alternatives. The average hourly payments (in NOK) of daycare alternatives were 29, 87, 87 and 67 for relatives, Nanny/Au-pair, formal childcare, and others respectively.

3. Data and econometric approach

The data are from national living standard surveys administered in the spring of 1998, 1999, and 2002, i.e. before and after the implementation of the CFC reform in the fall of 1998. These surveys collected information about the usage of different forms of childcare, as well as background characteristics of the families surveyed. Statistics Norway collected the data with the purpose of evaluating the effects of the reform.

We mainly concentrate on the following question asked in the surveys, “What form of care does your child have during daytime/working hours?” The question asked before and

[7] Eibak (2002) surveys the payment systems of 109 municipalities. 63 did not means test the formal childcare prize, 52 had 50% discount for the second sibling enrolled to childcare, and 23 had higher prices for children below age 3.

after the reform was in the form of a multiple-choice question, where the respondent indicated one or more care alternatives. The question in the 2002 survey was slightly different in that it comprised a separate question concerning parental care. To obtain a consistent measure of parental care across the three survey years, we coded "Parental care" for those respondents that did not identify any of the other care alternatives in all years. The choice of how to code parental care does not matter for the results. Both ordinary childcare centers and family childcare groups are defined as formal childcare.

For the 1998 survey, 2,500 mothers with at least one child born after 1.1.1992 were drawn. In addition, one thousand mothers with at least one child born after 1.1.1996 were drawn. Thus, families with very young children were oversampled. The response rate for the 1998 survey was 84.9%. The 1999 survey was based on a sample of 2,257 families from the first survey. Additional mothers were drawn to get a self-weighting sample. In total, 3,848 women were drawn for the 1999 survey, of which 86.6% responded. For the 2002 survey 2,700 mothers with at least one child born after 1.1.1996 were drawn. In addition, 1,200 women who had at least one child born after 1.1.1999 were sampled. 86.8% of the 3,886 mothers with preschool-aged children for the 2002 survey responded.[8]

We start with a sample containing childcare information on the first- and second-born children of the respondents in the household survey. We choose to concentrate on the married/cohabiting households. To get a valid comparison group, we exclude children aged 3-5 with CFC eligible siblings. We then excluded information on those children aged under 1 year and older than 5 years, leaving us with a baseline sample of 6,751 children.

[Insert Table 1 here]

Descriptive statistics for the sample are reported in Table 1. The table provides the averages and proportions of the most important variables used in the analysis. Panel A details

[8] In the 1998 survey 70% answered a self-reported postal questionnaire, while 15% answered through a telephone interview. The 1999 and 2002 surveys were both conducted using telephone interviews.

descriptive information on the background variables. The parental income and education variables are based on self-reported income in the previous year and administrative information on highest completed level of education. Income is measured in NOK. We note that the parents of 1–2 year olds and those of the 3–5 year olds appear similar in terms of background characteristics. For the subgroup analysis, we should also note that the immigrant population represents a relatively small proportion of the sample. We therefore expect the estimates for this group to be somewhat noisy. Panel B details the proportion of children across the different types of care alternatives. The biggest difference between the care of 1–2 and 3–5 year olds is that many more of the older group is in childcare.[9] Panel C of Table 1 provides the formal childcare attendance rates by subgroups. Children are defined as low SES if the mother does not have a university/college degree, and the father earns less than the 25th percentile in the distribution of earnings among fathers in the sample. The child is grouped into the immigrant category if the mother has an immigrant background. These sample splits show that children with high-SES backgrounds and non-immigrant mothers are more likely to attend formal childcare.[10]

To analyze the effect of the CFC reform on formal childcare attendance, we could compare formal childcare rates before and after the reform for eligible children. However, there could be underlying trends in formal childcare attendance rates for 1–2 year olds that have little or nothing to do with the CFC reform. To overcome this, we compare the change in formal childcare rates for eligible children to the change in formal childcare rates for ineligible children. The difference in the change in childcare rates is then attributed to the reform. The following difference-in-difference (DID) approach is specified as:

[9] Appendix Figure A.1 show these numbers in a graph.

[10] The two surveys after the reform lacked information on gender, something that prevents us from exploring rates separately for girls and boys (see for instance Kottelenberg and Lehrer, 2017).

$$Y_{it} = \beta_1 + \beta_2 D_{it}^{1999} + \beta_3 D_{it}^{2002} + \beta_4 D_{it}^{1999} D_{it}^{age1-2} + \beta_5 D_{it}^{2002} D_{it}^{age1-2} + \eta \; Controls_{it} + \varepsilon_{it} \quad (1)$$

where subscript *i* index the individual family child and *t* indexes time. The dependent variable Y_{it} is a binary variable equal to one if the parents responded that the daytime caregiver for their child is a formal childcare center, and zero otherwise. $D_{\mathrm{it}}^{\mathrm{age1\text{-}2}}$ is a binary variable denoting child aged 1–2 years at the time of the survey (i.e. one for an eligible child and zero otherwise), while D_{it}^{1999} and D_{it}^{2002} are dummies for recordings in 1999 and 2002, respectively. When constructing the interaction terms with the survey years, D^{1999} and D^{2002}, we use the dummy $D_{\mathrm{it}}^{\mathrm{age1\text{-}2}}$, not single age dummies (which are in the vector of *Controls*). This provides us with the effect for all eligible children. $Controls_{\mathrm{it}}$ consists of a set of control variables. The controls include an immigrant dummy, and dummies for the mother's and father's educational level (lower secondary, upper secondary, college and university). To account for the fact that the surveys conditioned on child birthyear when drawing families, age dummies are included. The control vector also includes a set of regional dummies (Oslo (Oslo, and Akershus), east excluding Oslo/Akershus (Hedmark, Oppland, Østfold, Vestfold, Buskerud and Telemark), southwest (Vest-Agder, Aust-Agder and Rogaland), west (Hordaland, Sogn og Fjordane, and M.-Romsdal), middle (Sør-Tr., and Nord-Tr.), and north (Nordland, Troms, and Finnmark)). Finally, $\varepsilon_{\mathrm{it}}$ is the error term. To correct for intragroup correlation in the error terms, standard errors are clustered at region-age level.

The parameters of main interest are β_4 and β_5, where β_4 captures the effect the year after the reform was implemented, while β_5, captures the effect four years after the reform was fully implemented. There could be different sources behind finding different effects one and three years after the CFC reform. The first one is based on the idea that the knowledge about the possibility and the generosity of the CFC benefits spread out over time and therefore affected the utilization of the CFC benefits.[11] However, we should note that the introduction of the CFC benefits was already described in the contract between the parties of the Christian-Green-Liberal minority collision government, ten months prior to the reform was effective~~.~~ The second source might be the difference in the level of benefit, being 2,263 NOK per month in 1999, and 3,000 NOK per month in 2002. A third source for different effects in 1999 and 2002, is supply side adjustments. The increase in childcare attendance for the control group shown in Table 1 provide evidence of this. The presence of excess demand could give a smaller response to the CFC reform than in a situation without excess demand, since newly freed up slots would be filled with other families not served by the market earlier (Gustafsson and Stafford, 1992). With this market situation, it is not obvious to what extent supply side adjustments will affect the relative size of estimates of childcare usage in 1999 and 2002. In total, delayed information spread and change in benefit size suggest larger effects in 2002 compared to 1999, while the market situation with excess demand suggest smaller effects for both coefficient than in a situation without.

4. Results

With the numbers already reported in Table 1 – Panel B Care alternatives "Formal childcare" – we calculate a first basic difference-in-difference estimate without controls of –0.096 for 1999 and –0.126 for 2002. These two estimates indicate that the formal childcare utilization for the eligible group dropped due to the introduction of the CFC reform.

The main results of our DID analyses are reported in Table 2. The table provides estimates of the parameters β_4 and β_5 in equation (1), i.e., the effects in 1999 and in 2002. Starting with the results in Column (1), which includes all children, we obtain a negative significant coefficient of –13.8 percentage points in 1999, and a negative and statistically

[11] See for instance Dahl *et al.* (2014) for similar findings for the introduction of paid paternity leave in Norway, and Rege *et al.* (2012) for the disability pension participation locally among older workers.

significant coefficient of –14.4 percentage points in 2002. The difference between the two coefficients is statistically insignificant (*p*-value = 0.81).

[Insert Table 2 here]

In Column (2), we concentrate on the youngest age group and therefore exclude 2-year-old children from the treatment group. Similarly, in Column (3), we focus on those 2-year-old children and exclude 1-year-old children. Comparing the results in Columns (2) and (3), we can see the estimates for younger children are larger in 1999 and in 2002. In 2002, the difference is significant at the 10% significance level.

Columns (4)-(15) of Table 2 provide estimates of the coefficients for the SES and immigrant status subgroups. The motivation behind this is to explore whether particular subgroups are more or less sensitive to the CFC reform. Columns (4)-(6) detail the estimates for the low-SES children. The point estimates are of a larger magnitude for low SES children than high SES children, reported in Columns (7)-(10). There is an argument that the CFC benefit redistributes to low-SES households since a lower proportion of children in these households attend childcare. Childcare is an in-kind public good that affects the distribution of (extended) income in the population (Aaberge and Langøren, 2006). The total distributional effect of the CFC reform must take into account the direct redistribution, the effect on parental labor force participation, and the effect of change in childcare use.[12]

In Columns (10)-(15), the sample is stratified by mother's immigration status. Since most of the sample consists of non-immigrants, the results for this group are almost identical to the main results. The results for immigrants are shown in Columns (10)-(12). They show a large and significant (at the 10% level) 2002 coefficient for 1-year-old immigrant children, while the remaining coefficients are not significantly different from zero. The positive, but

[12] Another perspective on distributional effects is how the CFC affects well-being across the income distribution (Burton and Phipps, 2007). Both income and parental time may affect family well-being, and low-SES families may gain more in terms of these factors. Furthermore, the CFC benefit is a transfer to family households. The benefit will the increase income inequalities between men with and without children (see for instance Kunze, 2016).

small coefficients for 2-year old immigrant children, Column (12), are the largest deviations from the other results reported in the table. Overall, we cannot conclude that the response to the price change is different for immigrants compared to non-immigrants.[13]

[Insert Table 3 here]

Table 3 details estimates of β_4 and β_5, the 2002 and 1999 effects, when the dependent variables are indicators of different forms of care. Column (1) provides estimates of the effect on all eligible children; Column (2) shows the effect only for 1-year-old children, while Column (3) shows the effect only for 2-year-old children. The 2002 effects, shown in Column (1), indicate that "Parental care" use increases by 9.4 percentage points and "Nanny" use increases by 3.6 percentage points, while "Other" use increases care attendance by around 4.6 percentage points. This latter alternative includes day parks, au pairs and uncategorized care alternatives. Looking at the 1999 effect, the main difference relative to the 2002 effect, is that "Nanny" use does not increase significantly. In total, both the 1999 and 2002 effects reported in Column (1) suggest that parental care is the most important alternative mode of care.

An interesting observation is that even though "Relative care" is an important care alternative for children, there is no observed increase in this type of care arrangement following the reform. There are several possible explanations. It could be that relative care is a complement to formal childcare. An alternative explanation is that relative care is an inferior good.

Columns (4)–(6) provide the results for low-SES children, while Columns (7)–(9) show the results for high-SES children. Comparing the results for the subgroups shows that the main alternative for formal childcare is parental care for low-SES children, while the alternative forms of care are more mixed for high-SES children. Nanny care and alternatives included in the "Other" category are both important alternatives for the latter group. We also

[13] Formal testing shows that we cannot reject the null hypothesis that the set of coefficients in Column (10) and the one in Column (13) are the same.

note that a significant negative 1999 coefficient shows up for "Relatives" for the high-SES group, suggesting that it is either a complement of formal childcare or and inferior good for this group. In contrast, the 1999 effects show an increase in the use of relative care for 1-year-olds of low SES parents.

Table 2, Column 1, also shows the 1999 and 2002 price elasticities of childcare (confidence intervals in square brackets). We use the estimated coefficients, average childcare expenses from our data in 1999 and 2002, and predicted counterfactual formal childcare attendance when calculating these elasticities.[14] The estimates are based on households where the mother has completed high school as the highest education in Oslo/Akershus. An observation is that the point estimates of the elasticities for 1 year olds are much larger than the estimates for 2 year olds. In contrast to the pattern observed for the coefficients, the point estimates of 1999 elasticities are larger than the 2002 elasticities. The 1999 overall elasticity is –0.33 [–0.49,–0.17], while the 2002 elasticity is –0.25 [–0.41,–0.09].[15]

It is useful to compare our measurement of the price sensitivity of formal childcare to other causal estimates in the literature. Baker *et al.* (2008) obtains a price elasticity of –0.58 for 0–4 year olds in Canada. This is of much larger magnitude (in absolute value) than our estimate, and we suspect part of the reason is that the Canadian reform included additional measures aimed at increasing the use of childcare other than the introduction of a subsidy. Gathmann and Sass (2017) estimate an elasticity of –0.60, which is also much larger than our estimate. Their estimates of the impacts on informal childcare alternatives also differ somewhat from our. They also find that mainly parental care increases with the CFC, but find

[14] We have estimated common counterfactual childcare attendance rates for 1-2 year olds using inverse probability age weights. We use observed childcare expenses for 1-2 year olds in 1999 and 2002 reported in the household surveys.

[15] The 2002 overall elasticity is calculated using the 2002 coefficient –0.136 (from a weighted regression), the predicted counterfactual 2002 formal childcare attendance rate of 0.495, the size of the CFC benefit of 3,000 NOK and the post-reform average payment for formal childcare given in the survey of 2,747 NOK. (–0.136 / 0.495) / (3,000 / 2,747) = –0.251. Note that the confidence intervals of elasticities do not take into account uncertainties in this measure of the price.

no impact on the use of nannies or "child-minders", while they do find a strong negative effect for care provided by friends/relatives. One possible reason for the larger effect found is that they report excess capacity in the relevant state at the time of the reform. The different estimates could also be explained by cross-country differences in childcare systems, or non-linear effects. Lastly, using Norwegian data, Kornstad and Thoresen (2007) estimate an elasticity for childcare of –0.12 for preschool-aged children (1–6 years) while Black *et al.* (2014) are unable to reject the null hypothesis of no effect on childcare use for the childcare subsidy for 5-year-old children. Such inelastic demand for childcare for groups of older children is consistent with our finding that the demand for formal childcare for younger children is more elastic than that for older children.

5. Robustness

Identification relies on a common trend in childcare rates over time for 1–2 year olds and 3–5 year olds. Unfortunately, it is not possible to derive earlier trends because the first survey including these questions was only conducted in 1998. However, there are official statistics on children in childcare by age for the total population from 1990 to 2003 (Statistics Norway 2005), as illustrated in Figure 1. These numbers are based on childcare centers reporting the number of children in care to Statistics Norway for general administrative purposes.

[Insert Figure 1 here]

The trends in formal childcare attendance for 1–2 and 3–5 year olds move very closely together before the reform.[16] Childcare rates for 1–2 year olds then increase on average 3.5 percentage points each year from 1990–1997, while those for 3–5 year olds increase at 3.2

[16] Rates are given in Appendix Table A.1. Note that the observed common pre-trends reported in Figure 1 are for formal childcare rates. Preferably, we would have like to have similar rates for the use of parental, relative or nanny care, but such information is not available.

percentage points each year for the same period.[17] There seems to be an increase in coverage rate in 1997. Average yearly change in coverage rate from 1990 to 1996 is 2.9 and 3.1 percentage points for children age 3-5 and 1-2 respectively. The corresponding increase in 1997 is 4.7 and 6.0 percentage points.[18] The possible existence of capacity constraints in the pre reform period, followed by the school starting-age reform, Reform 97, that excluded a full cohort of children from childcare from August 1997 onwards, is the likely explanation. Since there is very little grade retention in Norway, and enrollment follows the birth year, a large number of 6 year olds were no longer attending childcare by the end of 1997.[19] The increase then reflects excess demand being met by capacity made available by the school starting-age reform. Finally, the increase in the attendance rate tapers off from 1997 to 2000 for 1–2 year olds.[20]

Figure 1 also suggests that the childcare growth rate for 3–5 year olds slows somewhat after the CFC reform. The yearly increase before the reform was 3.2 percentage points, while after the reform it was 2.1 percentage points.[21] There are three potential explanations for this. First, children with siblings in childcare age could be affected by the reform.[22] Since we exclude this group from the treatment group, our main estimate based on the household surveys are unaffected by this. Second, there could be long-term effects of the CFC reform on the children affected. Children aged 3-5 after the reform could have been affected by the CFC

[17] A *t*-test on the difference in mean yearly changes in the pre-reform period between the two groups shows that they are not statistically different. Inference based on this test is valid if we assume that the yearly rate changes are independent.

[18] Testing for a significant different yearly change in coverage rate in 1997 using a regression in the pre-reform period shows that it is statistically larger.

[19] Note that only about 40% of 6 year olds were enrolled in regular formal childcare in 1996 because of a voluntary school preparation program (Norwegian Ministry of Children and Equality, 1991).

[20] From 1999, "open childcare" providers were no longer included as attending childcare. In 1998, children in these centers represented about 2% of all children in childcare (Statistics Norway 2005).

[21] Testing for the difference in mean yearly changes in the pre- and post-reform periods using a regression shows that they are statistically different.

[22] Appendix Table A.2 shows estimates of the effect on children aged 3-5 with siblings aged 1-2. Results are consistent with Bettinger *et al*. (2014) showing that older, ineligible siblings care type are affected by the CFC reform.

reform since they were eligible when they were younger. We assess that any bias arising for this source should be small. In any case, this should not affect the short-term coefficients since it is measured shortly after the implementation of the reform. The third explanation is that the 3-5 year olds could be affected by the newly freed up capacity in childcare following the CFC reform. The household surveys indicate this since we observe a large increase in attendance rate for 3-5 year olds from 1998 to 1999 in Table 1. Work- and rental contracts may still bind the childcare centers in 1999, so that they do not change the number of seats supplied. These restrictions may not be the same in 2002, leading to a delayed adjustment to the reform. If the demand shortfall for 1-2 year olds following from the CFC reform affects the childcare centers' response for both 1-2 year olds and 3-5 year olds in the same way, this does not affect the interpretation of our estimates.

The observed common pre trends are our main identification-test. Another way to explore whether the results reported in Table 2 are driven by factors other than the CFC reform is to run some regressions, as shown in equation (2), where A_{it} denotes a vector of age dummies. Differential trends in the observable variables would indicate that something other than the reform could explain the change in attendance rates.

$$Background_{it} = \alpha_1 + \alpha_2 D_{it}^{1999} + \alpha_3 D_{it}^{2002} + \alpha_4 D_{it}^{1999} D_{it}^{age1-2} + \alpha_5 D_{it}^{2002} D_{it}^{age1-2} + \kappa \mathrm{A}_{it} + \eta_{it} \quad (2)$$

[Insert Table 4 here]

Table 4 provides the results. The only significant difference is the likelihood of mother attending university, with six percentage points more university-educated mothers found among 1–2 year olds than 3–5 year olds in 1999 than in 1998 in our sample. This latter finding may point to a possible explanation for finding different effects in 1999 and 2002, and suggests that we should be somewhat more cautious when interpreting the 1999 effects.

As an additional robustness, we split the sample according to regional characteristics based on municipality specific information aggregated up to the six regions in our survey sample. These characteristics are such as unemployment rate, sickness absence rate, female employment level, and urbanization level, in addition to living in the capital area Oslo/Akershus. When analyzing subsamples, we find no significant differences between the regions split using these measures.[23] Since the subsamples are very equal on characteristics available at the regional, we cannot rule out that these characteristics can disguise interesting heterogeneity in effects based on other observables than those reported in Table (2).

Our study may be invalidated because not all the randomly drawn mothers responded to the survey and this may relate to the extent families were affected by the reform. For example, if high-SES mothers were more likely to respond to the survey, and these mothers were less affected by the subsidy, it could bias the estimated effect of the CFC benefit downwards. Further, if immigrant mothers who were more proficient in Norwegian were more likely to respond to the survey and less likely affected by the CFC subsidy, this would also weaken the results for this group. Since our results do not suggest that the effects are different for various subgroups, and the response rate was high, we do not consider this to be an important issue for our main estimates.

We are fortunate to have childcare attendance measurements from two sources: administrative childcare rates and household surveys. Thus, one way to validate our results from the surveys is to compare them to estimates based on administrative data. Figure 1 and Appendix Table A.1 detail the administrative rates. As the administrative data are reported at the end of year, we use the pre-reform rate from 1997, as the reform was implemented in August 1998. The surveys were conducted during spring each year. The 1999 and 2002 DID coefficients using the administrative data yield estimates of the effects of –6.5 and –9.3 percentage points, respectively.[24] Both of these estimates are admittedly outside the 95%

[23] These results are not reported, but available from the authors on request.

[24] The calculations are based on administrative numbers reported in Table A.1 recorded closest in time to the surveys (at the end of 1997, 1999 and 2001), such that the 1999 effect is (0.781–0.742) – (0.374–0.400) = –0.065, and the 2002 effect (0.815–0.742) – (0.380–0.400) = –0.093.

confidence interval for the estimates from the survey. This might question the external validity of our survey-based results. One explanation for the different estimates using administrative data and the survey data is likely that the populations in the two datasets are defined differently (see our Section 3 for details about defining our final survey sample). For example, 3-5 year olds with CFC eligible siblings are excluded from the comparison group in the survey, while they are included in the comparison group in the administrative data. Since Table A.2 shows that this group reduces attendance after the reform, this is a highly likely explanation for finding smaller effects in the administrative data. In addition, it could also be due to slight different definitions of formal childcare attendance in the survey and administrative data. Thus, even though the magnitudes of the results in the survey and administrative data differ somewhat, the main tendency is similar, and we therefore believe our results to be representative for the whole population.

Assuming formal childcare is a normal good, the substitution and income effect should decrease the demand in formal childcare following a regular price increase in formal childcare. The introduction of the CFC-benefit corresponds to a subsidy of non-publicly funded childcare. The substitution effect is expected to decrease demand for formal childcare, while the income effect works in the opposite direction. Thus, an income effect following the introduction of the CFC reform would work in the opposite direction compared to a regular price increase. Depending on the magnitude of the income effects, our estimate for formal childcare will be smaller than the effect of an ordinary price increase in formal childcare. If the other care alternatives are normal goods too, the substitution and income effect should both shift demand in the same direction for these alternatives.

6. Conclusion

This study focuses on the effect of a price change on childcare attendance. Few studies have isolated large shocks to childcare prices unaccompanied by other changes and studied the consequences in detail. The introduction of the CFC in Norway therefore provides a good opportunity to examine the response to price changes in formal childcare. Recent emphasis of the importance of childcare for later life outcomes motivates our main contribution of adding evidence on what the alternative mode of care to formal childcare is.

The price change accounted for a decrease in childcare attendance of 14.4 percentage points 4 years after the reform and mainly increased parental care. The magnitude of this effect implies price elasticity for formal childcare of about –0.25. Given the number of studies analyzing the consequences of childcare on children's future outcomes, it is important to have detailed information on what the alternative modes of care is. We find that the most important alternative to formal childcare in this context is parental care. This is in contrast to previous studies that have found informal childcare to be dominant alternative mode of care. Of course this could be different in different countries, and depend on the compliant group to the reform. That a nanny care and day park also is an alternative to formal childcare suggests that it attenuates the effect of providing affordable formal childcare on parental labor market participation. Our results suggest that formal childcare crowds out other out-of-home care alternatives, but not fully. Furthermore, the results show that out-of-home non-relative care is important for high-SES families but not low-SES families. Thus, the study demonstrates that the alternative form of care depends on important observable characteristics of the family.

Further studies should aim to explore the responses of different subgroups, such as different groups of immigrants, female/males and employed/unemployed, more so than has been possible with the data used in this analysis. Moreover, how the price sensitivity varies with different levels of excess demand would also enhance our understanding. Studies should also seek to explore whether the alternative mode of care are depend on whether the childcare decision is influenced by capacity expansions or price changes.

References

Aaberge, R. & Langørgen, A. (2006), Measuring the benefits from public services: The effects of local government spending on the distribution of income in Norway. *Review of Income and Wealth* 52(1), 61–83.

Baker, M., Gruber, J. & Milligan, K. (2008), Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy* 116(4), 709–745.

Bakken, F. & Myklebø, S. (2010), Kontantstøttens utbredelse og foreldres preferanser for barnetilsyn. NAV Rapport nr. 1.

Baklien, B., Ellingsæter, A. L. & Gulbrandsen, L. (2001), Evaluering av kontantstøtteordningen, Norges forskningsråd: Området for kultur og samfunn.

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

Bettendorf, L. J., Jongen, E. L. & Muller, P. (2015), Childcare subsidies and labour supply - evidence from a large Dutch reform. *Labour Economics* 36, 112–123.

Bettinger, E., Hægeland, T. & Rege, M. (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., Devereux, P. J., Løken, K. V. & Salvanes, K. G. (2014), Care or cash? The effect of child care subsidies on student performance. *Review of Economics and Statistics* 96(5), 824–837.

Blau, D. M. & Hagy, A. P. (1998), The demand for quality in child care. *Journal of Political Economy* 106(1), 104–146.

Blau, D. M. & Robins, P. K. (1988), Child-care costs and family labor supply. *The Review of Economics and Statistics* 70(3), 374–381.

Blix, K. (1993), Barnehage: Behov, etterspørsel og fordeling. rapport nr 93:8., Technical report, Institute for Social Research.

Blix, K. & Gulbrandsen, L. (1992), Norske familiers økonomiske levekår. første rapport fra en intervjuundersøkelse høsten 1991. Notat nr. 3. Institute for Social Research.

Burton, P. & Phipps, S. (2007), Families, time and money in Canada, Germany, Sweden, the United Kingdom and the United States. *Review of Income and Wealth* 53(3), 460–483.

Connelly, R. & Kimmel, J. (2003), Marital status and full–time/part–time work status in child care choices. *Applied Economics* 35(7), 761–777.

Dahl, G. B., Løken, K. V. & Mogstad, M. (2014), Peer effects in program participation', *The American Econmic Review* 104(7), 2049–2074.

Drange, N. (2015), Crowding out dad? The effect of a cash-for-care subsidy on family time allocation. *Nordic Journal of Political Economy* 40(2) 1-29.

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

Eibak, E. E. (2002), Survey on childcare prices 2002. Statistics Norway notater 2002/29.

Finseraas, H., Hardoy, I. & Schøne, P. (2017), School enrolment and mothers' labor supply: evidence from a regression discontinuity approach. *Review of Economics of the Household* 15(2), 1–18.

Gathmann, C. & Sass, B. (2017), Taxing childcare: Effects on family labor supply and children. *Journal of Labor Economics*, forthcoming

Gustafsson, S. & Stafford, F. (1992), Child care subsidies and labor supply in Sweden', *Journal of Human Resources* 27(1), 204–230.

Hardoy, I. & Schøne, P. (2010), Incentives to work? The impact of a cash-for-care benefit for immigrant and native mothers labour market participation. *Labour Economics* 17(6), 963–974.

Hardoy, I. & Schøne, P. (2015), Enticing even higher female labor supply: The impact of cheaper day care. *Review of Economics of the Household* 13(4), 815–836.

Havnes, T. & Mogstad, M. (2011), Money for nothing? Universal child care and maternal employment. *Journal of Public Economics* 95(11), 1455–1465.

Heckman, J. J. & Mosso, S. (2014), The economics of human development and social mobility. *Annual Review of Economics* 6, 689–733.

Kalb, G. & Thoresen, T. O. (2010), A comparison of family policy designs of Australia and Norway using microsimulation models. *Review of Economics of the Household* 8(2), 255–287.

Kornstad, T. & Thoresen, T. O. (2007), A discrete choice model for labor supply and childcare. *Journal of Population Economics* 20(4), 781–803.

Kottelenberg M. J & Lehrer, S. F. (2017), Does Quebec's Subsidized Child Care Policy Give Boys and Girls an Equal Start?. *NBER Working Paper* No. 23259.

Kunze, A. (2016), The effect of children on earnings inequality among men. *IZA Discussion Paper* No. 8813.

Lefebvre, P. & Merrigan, P. (2008), 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.

Leibowitz, A., Klerman, J. A. & Waite, L. J. (1992), Employment of new mothers and child care choice: Differences by children's age. *Journal of Human Resources* 27(1), 12–133.

Lundholm, M. & Ohlsson, H. (1998), Wages, taxes and publicly provided day care', *Journal of Population Economics* 11(2), 185–204.

Lundin, D., Mörk, E. & Öckert, B. (2008), How far can reduced childcare prices push female labour supply? *Labour Economics* 15(4), 647–659.

Naz, G. (2004), The impact of cash-benefit reform on parents labour force participation, *Journal of Population Economics* 17(2), 369–383.

Norwegian Ministry of Children and Equality (1991), Ot.prp. nr. 57 (1990-1991): Om lov om endring av lov 6.juni 1975 nr. 30 om barnehage m.v.

Norwegian Ministry of Education (1996), Reform 97 - Dette er grunnskolereformen., *Government white paper*

NSD (2017), Kommunedatabasen. Norwegian Centre for Research Data – NSD.

OECD (2016), Organisation for Economic Co-operation and Development Family Database, OECD Paris, France.

Powell, L. M. (2002), Joint labor supply and childcare choice decisions of married mothers. *Journal of Human Resources* 37(1), 106–128.

Rege, M., Telle, K. & Votruba, M. (2012), Social interaction effects in disability pension participation: evidence from plant downsizing. *The Scandinavian journal of Economics* 114(4), 1208–1239.

Reppen, H. K. & Rønning, E. (1999), Barnefamiliers tilsyns- ordninger, yrkesdeltakelse og bruk av kontantstotte varen 1999, SSB Reports 1999/27 .

Ribar, D. C. (1992), Child care and the labor supply of married women: Reduced form evidence. *Journal of Human Resources* 27(1) , 134–165.

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

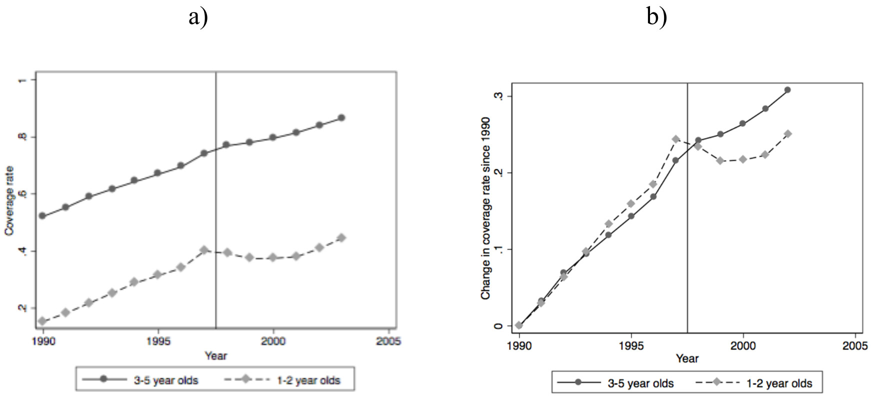
Statistics Norway (2005), Official Statistics of Norway: Kindergartens 2003 NOS D 328.

Statistics Norway (2017), Labor Force Survey 1972-2017.

Tekin, E. (2005), Child care subsidy receipt, employment, and child care choices of single mothers. *Economics Letters* 89(1), 1–6.

Tekin, E. (2007), Childcare subsidies, wages, and employment of single mothers. *Journal of Human Resources* 42(2), 453–487.

Figure 1: Formal childcare rates



Notes: (a) End of year childcare coverage rates as reported in the official statistics based on yearly status reports from childcare providers and sent to Statistics Norway. (b) The percentage point change since 1990 (Statistics Norway 2005)

Table 1: Descriptives

	1-2 year olds				3-5 year olds			
	1998	1999	2002	diff1998-2002	1998	1999	2002	diff1998-2002
Panel A - Background								
Mother income	155.7	179.3	210.8	55.1	158.6	186.8	226.5	67.8
Father income	269.5	287.0	287.3	17.8	286.7	313.0	318.8	32.1
Immigrant	0.056	0.062	0.073	0.017	0.062	0.064	0.072	0.010
Low-SES	0.178	0.207	0.154	-0.024	0.176	0.164	0.142	-0.034
Oslo/Akershus	0.193	0.213	0.207	0.014	0.199	0.194	0.198	-0.001
Mother education								
Lower Secondary	0.045	0.048	0.057	0.012	0.068	0.064	0.057	-0.011
Upper Secondary	0.539	0.541	0.514	-0.026	0.527	0.592	0.558	0.031
University	0.376	0.373	0.411	0.035	0.370	0.312	0.371	0.001
Missing	0.040	0.038	0.019	-0.021	0.035	0.032	0.014	-0.021
Panel B - Care alternatives								
Formal childcare	0.356	0.327	0.335	-0.021	0.661	0.728	0.766	0.105
Parental care	0.314	0.322	0.330	0.016	0.173	0.115	0.105	-0.069
Relatives	0.180	0.149	0.144	-0.036	0.081	0.081	0.067	-0.015
Nannies	0.178	0.186	0.184	0.006	0.077	0.076	0.055	-0.022
Other	0.051	0.059	0.062	0.011	0.107	0.081	0.069	-0.038
N	1168	1621	1376		738	904	944	
Panel C - Formal childcare attendance by group (socioeconomic- and immigrant status)								
Low-SES	0.226	0.226	0.203	-0.023	0.523	0.676	0.716	0.193
High-SES	0.384	0.353	0.359	-0.025	0.691	0.738	0.774	0.083
Immigrant	0.268	0.296	0.278	0.011	0.578	0.729	0.711	0.132
Non-immigrant	0.366	0.331	0.342	-0.023	0.672	0.728	0.774	0.102

Notes: Parental income (measured in 1,000 NOK) and education variables are based on self-reported values the previous year. Childcare measures are based on responses to a survey of main daytime care of the respondents' children. Formal childcare includes municipal, private, and family childcare centers. Children are defined as low-SES if the mother do not have a university/ college degree and the father earned less than the 25th income percentile. Pre-reform children are from surveys conducted before the reform during spring 1998, while post-reform children are from surveys conducted after the reform during spring 1999 and 2002.

Table 2: Formal childcare attendance

	Overall			Low SES			High SES			Immigrants			Non_immigrants		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	all	age 1	age2	all	age 1	age2	all	age 1	age2	all	age 1	age2	all	age 1	age2
1999 effect	-0.138***	-0.160***	-0.104***	-0.184**	-0.188**	-0.170	-0.129***	-0.157***	-0.090***	-0.144	-0.246	0.011	-0.137***	-0.155***	-0.107***
	(0.032)	(0.043)	(0.036)	(0.083)	(0.090)	(0.101)	(0.033)	(0.044)	(0.032)	(0.148)	(0.135)	(0.182)	(0.031)	(0.043)	(0.033)
2002 effect	-0.144***	-0.172***	-0.103**	-0.225***	-0.283***	-0.161*	-0.133***	-0.158***	-0.095**	-0.103	-0.207*	0.043	-0.149***	-0.173***	-0.111**
	(0.026)	(0.025)	(0.041)	(0.076)	(0.078)	(0.089)	(0.029)	(0.028)	(0.045)	(0.115)	(0.121)	(0.132)	(0.028)	(0.028)	(0.044)
N	6751	4789	4548	1168	814	766	5583	3975	3782	437	317	292	6314	4472	4256
p-value (1999 = 2002)	0.81	0.72	0.97	0.59	0.19	0.94	0.89	0.98	0.88	0.63	0.71	0.75	0.65	0.57	0.91
p-value (age1 = age2)															
(in 1999)		0.24			0.87			0.13			0.07			0.28	
(in 2002)		0.10			0.14			0.16			0.01			0.16	
1999 elasticity	-0.33	-0.47	-0.22	-0.41	-0.48	-0.32	-0.30	-0.46	-0.19	-0.37	-0.68	0.02	-0.33	-0.45	-0.23
95% CI	[-0.49, -0.17]	[-0.64, -0.29]	[-0.34, -0.09]	[-0.65, -0.17]	[-0.74, -0.21]	[-0.62, -0.03]	[-0.47, -0.14]	[-0.65, -0.27]	[-0.30, -0.07]	[-0.89, 0.14]	[-1.11, -0.24]	[-0.79, 0.83]	[-0.48, -0.17]	[-0.63, -0.28]	[-0.34, -0.11]
2002 elasticity	-0.25	-0.38	-0.16	-0.31	-0.45	-0.20	-0.23	-0.35	-0.15	-0.27	-0.76	0.13	-0.26	-0.37	-0.17
95% CI	[-0.41, -0.09]	[-0.55, -0.20]	[-0.29, -0.04]	[-0.55, -0.07]	[-0.71, -0.19]	[-0.49, 0.09]	[-0.40, -0.07]	[-0.54, -0.16]	[-0.27, -0.04]	[-0.79, 0.25]	[-1.20, -0.32]	[-0.68, 0.94]	[-0.41, -0.10]	[-0.54, -0.20]	[-0.29, -0.06]

Notes: Estimates of the 1999 effect and 2002 effects on formal childcare attendance are shown. Column (1) includes all treated children, while Column (2) includes 1-year-old children and Column (3) includes 2-year-old children. The same estimations are performed on subgroups indicated in table header from Column (4) to Column (15). Control variables are education level, region dummies, immigrant status where applicable, and age dummies. P-values for equality of the 1999 and 2002 effects are shown. P-values of tests for equality of effects for children of different ages are shown. Standard errors clustered at region–age level and shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 3: Childcare alternatives

	All			Low-SES			High-SES		
	All	Age 1	Age 2	All	Age 1	Age 2	All	Age 1	Age 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Childcare									
1999 effect	-0.138***	-0.152***	-0.096**	-0.184**	-0.188**	-0.170	-0.129***	-0.157***	-0.090***
	(0.032)	(0.043)	(0.038)	(0.083)	(0.090)	(0.101)	(0.033)	(0.044)	(0.032)
2002 effect	-0.144***	-0.165***	-0.096**	-0.225***	-0.283***	-0.161*	-0.133***	-0.158***	-0.095**
	(0.026)	(0.028)	(0.043)	(0.076)	(0.078)	(0.089)	(0.029)	(0.028)	(0.045)
Parents									
1999 effect	0.094***	0.063**	0.097***	0.112	0.010	0.229**	0.089***	0.089***	0.076**
	(0.024)	(0.026)	(0.030)	(0.083)	(0.075)	(0.100)	(0.022)	(0.025)	(0.029)
2002 effect	0.094***	0.130***	0.035	0.179**	0.220***	0.156	0.080***	0.126***	0.019
	(0.024)	(0.025)	(0.027)	(0.065)	(0.074)	(0.100)	(0.025)	(0.024)	(0.028)
Relatives									
1999 effect	-0.018	0.004	-0.043*	0.050	0.135**	-0.063	-0.034*	-0.027	-0.037*
	(0.021)	(0.026)	(0.022)	(0.060)	(0.052)	(0.069)	(0.019)	(0.027)	(0.018)
2002 effect	-0.017	-0.021	-0.013	0.058	0.103	-0.020	-0.028	-0.045**	-0.007
	(0.021)	(0.025)	(0.023)	(0.063)	(0.064)	(0.063)	(0.020)	(0.021)	(0.023)
Nannies									
1999 effect	0.019	0.036	0.008	-0.005	0.013	-0.023	0.028	0.042	0.015
	(0.023)	(0.032)	(0.021)	(0.045)	(0.059)	(0.058)	(0.025)	(0.037)	(0.020)
2002 effect	0.036*	0.023	0.051	0.002	-0.020	0.018	0.042*	0.029	0.056
	(0.020)	(0.016)	(0.030)	(0.041)	(0.053)	(0.052)	(0.023)	(0.020)	(0.037)
Other									
1999 effect	0.027**	0.032**	0.022	-0.007	0.020	-0.046	0.033**	0.033*	0.035*
	(0.013)	(0.015)	(0.020)	(0.036)	(0.034)	(0.045)	(0.013)	(0.016)	(0.017)
2002 effect	0.046***	0.042**	0.045**	0.006	0.013	-0.011	0.052***	0.049***	0.056**
	(0.014)	(0.015)	(0.019)	(0.027)	(0.027)	(0.033)	(0.015)	(0.017)	(0.020)
N	6751	4789	4548	1168	814	766	5583	3975	3782

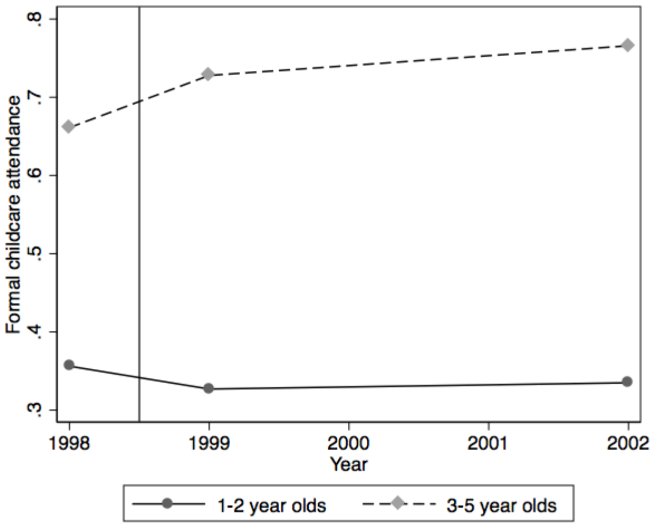
Notes: Estimates of the effect on care alternatives are shown. Dependent variable indicates attendance at care alternative. Column (1) includes all children, while Column (2) includes 1-year-old children, and Column (3) includes 2-year-old children. Columns (4)–(6) provides the results for low-SES children, while Columns (7)–(9) provide the results for high-SES children. Control variables are education level, region dummies, immigrant status and age dummies. Standard errors clustered at region–age level and shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 4: Robustness

	Mother income	Father income	Mother att. university	Immigrant	Oslo/Akershus
	(1)	(2)	(3)	(4)	(5)
1999 effect	-3.350	-9.940	0.060**	0.005	0.029
	(9.553)	(7.739)	(0.029)	(0.010)	(0.025)
2002 effect	-11.973	-15.005	0.037	0.008	0.017
	(9.852)	(12.838)	(0.028)	(0.014)	(0.022)
N	5574	6321	6751	6751	6751

Notes: Tests of different trends in background characteristics for the treatment and control groups are shown. The estimates of the 1999 effect and 2002 effect from equation (2) are shown. The dependent variables are the background characteristics of the parents of the children in the sample. Age dummies are included in all regressions. There are some missing observations for income variables. Standard errors clustered at region–age level and shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Appendix Figure A.1: Formal childcare attendance - Surveys



Notes: The figure shows formal childcare attendance rates given in the household surveys. Numbers are from based on descriptives reported in Table 1.

Table A.1: Attendance rates administrative data and household surveys

Year	Administrative*		Survey+		Survey+#	
	Age 1-2	Age 3-5	Age 1-2	Age 3-5	Age 1-2	Age 3-5
	(1)	(2)	(3)	(4)	(5)	(6)
1990	0.154	0.522				
1991	0.183	0.553				
1992	0.218	0.590				
1993	0.253	0.618				
1994	0.288	0.644				
1995	0.314	0.669				
1996	0.340	0.695				
1997	0.400	0.742				
1998	0.392	0.769	0.356	0.661	0.375	0.662
1999	0.374	0.781	0.326	0.730	0.320	0.737
2000	0.375	0.795				
2001	0.380	0.815				
2002	0.409	0.840	0.335	0.766	0.341	0.768
2003	0.446	0.866				

Notes: Estimated rates of children in childcare from 1990 to 2003 are shown. (*) End of year figures. (+) Spring figures (#) Weigted proportion - each observation weighted by the inverse of proportion of children of the same age. Source: Statistics Norway (2005).

Table A.2 - Effect of on older siblings

Dep. var: Care alternative	Formal childcar	Parental care	Relatives	Nannies	Other
	(1)	(2)	(3)	(4)	(5)
1999 effect	-0.108*** (0.029)	0.104*** (0.019)	-0.078*** (0.021)	-0.032 (0.020)	0.027 (0.020)
2002 effect	-0.072** (0.032)	0.080*** (0.025)	-0.076*** (0.023)	-0.022 (0.022)	0.017 (0.020)
N	4137	4137	4137	4137	4137

Notes: The table shows estimates of the 1999 effect and 2002 on care alternative of older ineligible siblings of eligible children. The dependent variable is an indicator for care type attendance. The treatment group are now children aged 3-5 with sibling aged 1-2 before and after the reform, while the control group are children aged 3-5 without siblings aged 1-2 before and after the reform. Control variables are education level, region dummies, immigrant status and age dummies. Standard errors clustered at region–age level and shown in parentheses. * p<0.10, ** p<0.05, *** p<0.01