

Child care for all? Treatment effects on test scores under essential heterogeneity*

Martin Eckhoff Andresen[†]

November 23, 2016

Abstract

The common claim that universal child care equalizes differences in endowments rests crucially on heterogeneity in the treatment effect. In this paper, I investigate how test scores in reading, math and English at age 10 are affected by early child care attendance using marginal treatment effects. Identification comes from a large Norwegian reform that expanded highly rationed child care for toddlers unequally across municipalities and over time. Results suggest that a) IV results are small and even negative, but hide important heterogeneity, b) there is some evidence of positive selection on observable gains and c) MTEs are downward sloping, suggesting selection on unobservable gains: Children with large unobserved treatment effects are most likely to be treated. This implies that the effect of further expansions is smaller than initial rollout as the children with the largest treatment effects are already served.

JEL-codes: J13, I26

Keywords: Universal child care, test scores, heterogeneity, marginal treatment effects

*The author wish to thank Edwin Leuven, Rafael Lalive, Thomas Cornelissen and Christian Brinch for advice and help at various stages of this project, as well as seminar participants at UiO, NHH and Statistics Norway. All errors remain my own. While carrying out this research I have been associated with the center for Equality, Social Organization, and Performance (ESOP) at the Department of Economics at the University of Oslo. ESOP is supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 179552.

[†]University of Oslo and Statistics Norway. E-mail to m.e.andresen@econ.uio.no

1 Introduction

Early child care programs targeted at disadvantaged children have a long tradition in the U.S. with programs such as Head Start and various small-scale interventions. These programs are generally considered relatively efficient, at least in the short run (Cascio and Schanzenbach, 2013; Almond and Currie, 2011; Fitzpatrick, 2008; Currie and Thomas, 1995), with substantial gains in cognitive test scores. Even though some studies document fadeout of these gains in the medium run, others document stable effects even after 30 years. The efficiency of these programs could be due to the simple fact that some skills are best learned young, due to the longer time period of pay-off when skills are acquired early or due to dynamic complementarities in learning (Cunha and Heckman, 2007) that allows early investment to increase the efficiency of later schooling. SHORTEN

In recent years, large-scale, universal child care programs have gained considerable attention from policymakers. In the U.S., President Obama has advocated the introduction of universal child care programs through his zero to five-plan, and policymakers from countries such as Korea, Australia and others are considering similar policies. While the traditional programs have been explicitly targeted at disadvantaged populations in order to allow them to catch up to their more advantaged peers, proponents of universal programs claim they have the same property: They equalize initial endowments by benefiting disadvantaged children more than advantaged children (Currie, 2001). One argument that is often made in favor of this is the differences in the alternative mode of care: Child care programs push disadvantaged children out of relatively low quality home environment into higher quality formal care, while this difference is smaller for advantaged children.

This argument crucially rests on heterogeneity in the treatment effect of child care programs: If all children benefit the same way, disadvantaged children cannot catch up when the same program is offered to everyone. Many papers have documented heterogeneity in who benefits based on observables, and often find that children from more disadvantaged populations benefit more. This literature considers heterogeneity in observables such as family background, gender and socioeconomic status

Another potentially important source of heterogeneity is preference for child care, but this is rarely observed. To address this issue, a few recent working papers use the marginal treatment effects framework to separate heterogeneity into the effect of observables and unobserved preferences on the treatment effects. Among these are Felfe and Lalive (2015) and Cornelissen et al. (2016), who both use expansions of universal child care programs in Germany to instrument for individual enrollment in marginal treatment effect programs and look at effects on school readiness. Felfe and Lalive (2015) look at child care for toddlers and find a downward-sloping MTE curve consistent with positive selection on unobserved gains: Parents with high preference for care also have high treatment effects. In contrast, Cornelissen et al. (2015) look at care for 3-6 year olds, and find reverse

selection on unobserved gains.

Both these papers restrict the municipality-, year- and birth month fixed effects to be the same in the treated and untreated state. In rationed markets that both these papers study, there are two-sided selection: Parents must apply for care, and they must be able to secure a slot under whatever rationing mechanism is in place. Without further analysis, we cannot know whether preferences or rationing are generating the shape of the MTE curve.

The universal welfare states of Scandinavia provide a natural case for evaluating the heterogeneity in universal child care programs. This paper builds on these previously mentioned papers by using a similar, massive expansion of child care for the youngest children in Norway to investigate the heterogeneity in treatment effects of early child care on test scores in the medium run. Following the Child Care Concord in 2002, child care availability expanded rapidly in the rationed market for child care, and I exploit the variation in this expansion between municipalities and over time. I relax the restriction that the fixed effects are the same in the two treatment states, and document how this is important in my setting. Using comprehensive data and a large-scale expansion of care, I generate almost full support of propensity scores, and can therefore evaluate the MTE over most of the unit interval. I investigate the effect of early child care attendance around age 2 on test scores in math, English and reading at age 10, complementing previous studies that look at school readiness. Using unique data on the number of children on waiting lists by municipality and year, I can also evaluate the degree of rationing.

Using the MTE framework, I document first that small and even negative traditional IV estimates hide important heterogeneity: They reflect local average treatment effects for very specific groups that turn out to have quite different treatment effects than the average. Second, I find a downward sloping MTE for math scores, while the MTEs for reading and English are relatively flat. This helps us reconcile the small, even negative IV estimates with substantial average gains in test scores. Third, I document the patterns of heterogeneity in treatment effects among observables, confirming that children from disadvantaged households have more to gain from child care along most dimensions. Lastly, I document how a continued expansion of child care is not likely to yield considerable gains by computing the policy-relevant treatment effect for a further expansion, finding that effects are low because the children with the highest gains are already served.

This paper proceeds as follows: In section 3 I briefly present the institutional setting and the child care reform. In section 2 I briefly present the MTE framework and the underlying choice model as well as the methods used to identify the MTE. Section 4 presents the empirical strategy, the instruments and the controls, while section 5 presents the data sources, important variable definitions and some descriptive statistics. Finally, section 6 presents the results and analyses, 7 a battery of specification checks and section 8 concludes.

2 Marginal treatment effects

To analyze the heterogeneity in the treatment effect, I rely on the marginal treatment effects framework. MTE is based on a generalized Roy model:

$$Y_j = X\beta_j + U_j \quad \text{for } j = 0, 1 \quad (1)$$

$$Y = DY_1 + (1 - D)Y_0 \quad (2)$$

$$D = \mathbb{1}[\gamma Z > V] \quad \text{where } Z = (X, Z_-) \quad (3)$$

Where Y_1 and Y_0 are the potential test scores in the treated and untreated state, respectively. They are both modeled as functions of observables X , where a subset of them are potentially restricted to have the same effect across treatment states.¹ Equation (3) is the selection equation, and can be interpreted as a latent index. It is thus a reduced form way of modeling the selection into treatment. Like the potential outcomes, the latent utility of treatment is modeled as a function of observables X and instruments Z_- that affect the probability of treatment, but not the potential outcomes.

The unobservable V in the choice equation is a negative shock to the latent index determining treatment. It is often interpreted as unobserved resistance to treatment. As long as the unobservable V has a continuous distribution, we can rewrite the selection equation as $P(Z) > U_D$, where U_D represents the percentiles of the unobserved resistance to treatment and $P(Z)$ the propensity score.² U_D , by construction, has a uniform distribution in the population.

The marginal treatment effects are the main parameters of interest in this paper. MTE was introduced by Björklund and Moffitt (1987), later generalized by Heckman and coauthors (1999; 2001; 2005; 2007), and is defined as

$$\begin{aligned} \text{MTE}(x, u) &\equiv \mathbb{E}(Y_1 - Y_0 | X = x, U_D = u) \\ &= x(\beta_1 - \beta_0) + \mathbb{E}(U_1 - U_0 | X = x, U_D = u) \end{aligned}$$

They measure test score gains for people with particular values of X and the unobserved resistance to treatment U_D . Alternatively, the MTE can be interpreted as the mean return to early care for children at a particular margin of indifference. Identification of this model requires the following assumptions:

¹In practice, this means setting $\beta_1 = \beta_0$ for a subset of X . Note that in the empirical section, X will contain year- and municipality fixed effects, and none of the X will be restricted. This possibility is included here to nest models used in other papers (Felfe and Lalive, 2015; Cornelissen et al., 2015), that restricts the fixed effects to have the same effect across treatment states.

²This is because $D = 1 \Leftrightarrow \gamma Z > V \Leftrightarrow F(\gamma Z) > F(V) \Leftrightarrow P(Z) > U_D$ where F is the cumulative distribution function of V .

Assumption 1: Exclusion (U_0, U_1, V) is independent of Z_- conditional on X

Assumption 2: Additive separability $E(Y_j|U_D, X) = X\beta_j + \mathbb{E}(U_j|U_D)$. The conditional expectation of U_j is independent of X .

Assumption 1 is nothing more than the standard exclusion restriction for IV. Vytlačil (2002) shows how the standard IV assumptions of relevance, exclusion and monotonicity³ (Imbens and Angrist, 1994) are equivalent to some representation of a choice equation as in (3). For the outcome equations, assumption 2 require additive separability of the conditional expectation of U_j from X . This implies that the shape of the MTE curve does not depend on X . Note how these assumptions are implied by, but does not imply, full independence between the errors and Z . This separability assumption is common in applied work using marginal treatment effects, see for example Carneiro et al. (2011); Felfe and Lalive (2015); Maestas et al. (2013); Carneiro and Lee (2009); Cornelissen et al. (2016).

Identification using local instrumental variables

Marginal treatment effects can be estimated using the method of local instrumental variables developed by Heckman and Vytlačil (1999, 2001, 2005). This method identifies the MTEs as the derivative of the conditional expectation of Y with respect to the propensity score:

$$\begin{aligned}\mathbb{E}(Y|X = x, P(Z) = p) &= \mathbb{E}[Y_0 + D(Y_1 - Y_0)|X = x, P(Z) = p] \\ &= x\beta_0 + x(\beta_0 - \beta_1)p + \underbrace{p\mathbb{E}[U_1 - U_0|U_D \leq p]}_{K(p)}\end{aligned}\quad (4)$$

$$\text{MTE}(x, u) = \frac{\partial \mathbb{E}(Y|X = x, P(Z) = p)}{\partial p} \Big|_{p=u} \quad (5)$$

$$= x(\beta_1 - \beta_0) + \underbrace{\mathbb{E}(U_1 - U_0|U_D = u)}_{k(u)} \quad (6)$$

Where I utilize that we can normalize $\mathbb{E}(U_0) = 0$ as long as X include an intercept. The MTE framework separates heterogeneity in the treatment effect into two parts: One part is heterogeneity in observables, the other heterogeneity in the unobserved resistance to treatment. $k(u)$ is a nonlinear function of u , and allows for nonlinearity of the treatment effect in the resistance to treatment. If people with different resistance to treatment have different expectations of their error terms in the treated and untreated state, the MTE will be nonlinear. The K notation follows Brinch et al. (2015).

³Or rather, uniformity, as it is a condition across people, not across realizations of Z : This assumption require that $\Pr(D = 1|Z = z) \geq \Pr(D = 1|Z = z')$ or the other way around for all people, but it doesn't require full monotonicity in the classical sense.

This decomposition suggest a simple estimation approach: First estimate the propensity scores p , and then estimate the conditional expectation of Y given X and p , using a parametric or semiparametric specification for $K(p)$. Taking the derivative of this expression at particular values of u and X yields the MTE estimates. For more details on this estimation procedure, see Andresen (2016).

Identification using the separate approach

Alternatively, as suggested by Heckman and Vytlačil (2007), MTEs can be estimated using the separate approach. This has the benefit of estimating all the parameters of the potential outcomes, so that we can plot these over the distribution of U_D . In practice, this means estimating the expectation of Y separately in the treated and untreated sample, controlling for selection:

$$\begin{aligned}\mathbb{E}(Y_1|X = x, D = 1) &= x\beta_1 + \mathbb{E}(U_1|U_D \leq p) = x\beta_1 + K_1(p) \\ \mathbb{E}(Y_0|X = x, D = 0) &= x\beta_0 + \mathbb{E}(U_0|U_D > p) = x\beta_0 + K_0(p)\end{aligned}$$

The MTE is then estimated as

$$\begin{aligned}\text{MTE}(x, u) &= \mathbb{E}(Y_1|X = x, U_D = u) - \mathbb{E}(Y_0|X = x, U_D = u) \\ &= (\beta_1 - \beta_0)x + k_1(u) - k_0(u) \\ \text{where } k_j(u) &= \mathbb{E}(U_j|U_D = u)\end{aligned}$$

In practice, the separate approach estimates the conditional expectation as in a selection model, where the terms $K_j(p)$ controls selection. In practice, first estimate the propensity scores p , and use these to construct $K_j(p)$ depending on the parametric or semiparametric specification. Then estimate the conditional expectation of Y separately in the treated and untreated sample.

3 Institutional setting and the child care reform

The Norwegian system of universal child care was introduced after WWII⁴ as a response to increasing female labor force participation and the goal of gender equality in the Nordic welfare model (Ministry of Education and Research, 1998). Increased excess demand for formal care in the 60's and 70's led to the Kindergarten Act of 1975, and a strong increase in the supply of formal child care for preschool children (Havnes and Mogstad,

⁴See Ministry of Education (2008-2009) for a thorough treatment.

2011), eventually leading to a high coverage rate for preschool children in 2000, as can be seen in figure 1d. At the same time, coverage was much lower for younger children. In 2001, less than 40% of 1–2 year olds in Norway were enrolled in child care, and there was substantial excess demand for child care.

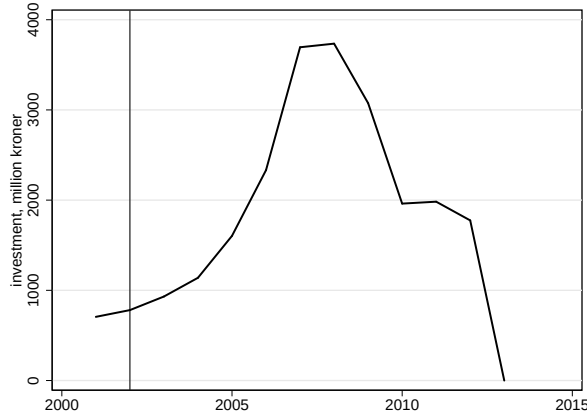
This excess demand was the background for the Kindergarten concord, a reform passed in the Norwegian Parliament in 2002 with broad bipartisan support. The main goal of the reform was to offer affordable care to all children, and to secure quality and diversity in child care services (Ministry of Education and Research, 2002-2003). The most important means for obtaining these goals were increased subsidies, lower parental fees and an investment subsidy for construction of new child care institutions. The concord also established equal treatment of private and public child care institutions, where private institutions had previously been awarded only 85% of the subsidy rate offered to public institutions, and thus made it easier for private suppliers of care to enter the market.

Figure 1 presents some important changes in the child care sector following the reform. Panel 1a depicts the total investment in child care institutions over the period.⁵ Note that most investments appear with a lag of 1–2 years. The figure shows clearly how total investments increased rapidly following the reform. Panel 1b shows the increase in total subsidies per child per year, and also shows steady increases over the period. Panel 1c shows the changes in the composition of the costs that was covered by the municipality, the central government and the parental fees. It is clear that the share of the costs covered by parents has declined significantly. This figure also shows that the municipal support was not reduced as a response to the increased government subsidies. The large overall increase in expenditures over the period is a result of both more children in care, higher subsidy rates, and an increasing share of toddlers, requiring more staff and resources per child.

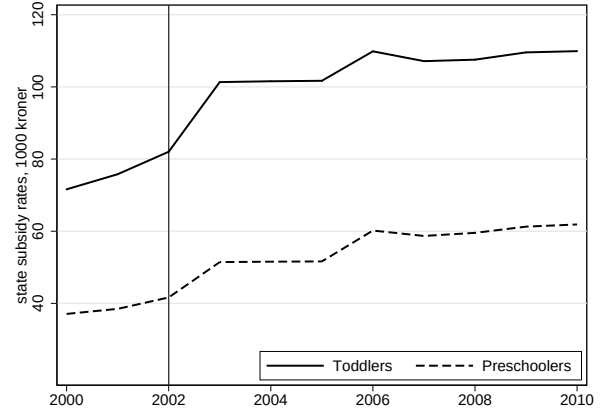
Throughout the period under study, formal childcare was highly regulated in Norway. In order to be eligible for the large subsidies offered following the reform, both private and public child care institutions have to adhere to strict quality criteria governing the quality of the services supplied. These criteria relate to for instance the teacher–child ratio, opening hours, and available playing area per child. Since 2004, the institutions also had to adhere to a max price reform that put a cap on the fee that can be charged from parents for a full time slot. In 2016 this cap stood at 2,655 NOK, or around 320 USD. In practice, this ensures that the formal child care institutions are relatively homogenous in terms of observable attributes of quality and price.

Figure 2 shows the aggregate application- and acceptance rates at the municipality level over time, constructed from data on the number of kids in care and on waiting lists for care. Firstly, notice how average demand increased rapidly over the period, but despite this, acceptance rates did not drop. This is a result of the large increase in availability of

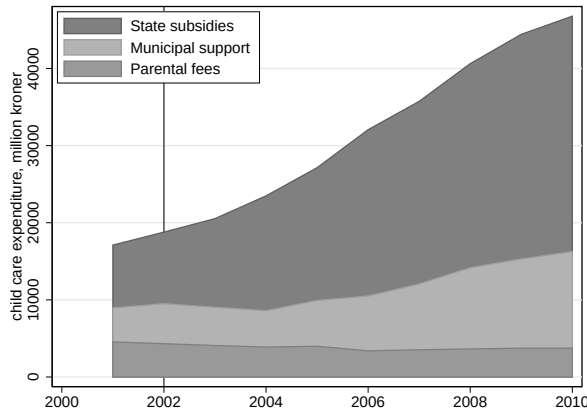
⁵All monetary values are indexed to 2014-kroners.



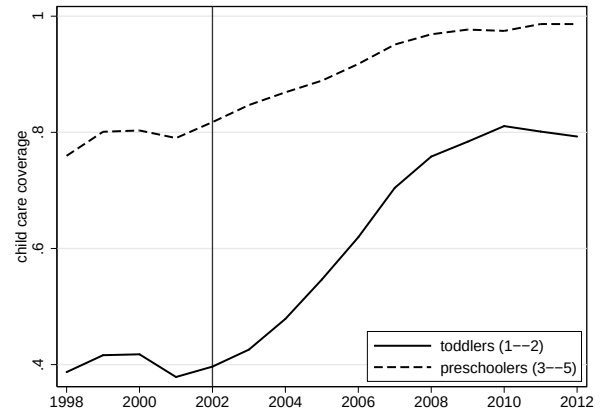
(a) Investment



(b) Yearly state subsidies per child



(c) The composition of financing in child care



(d) Child care coverage rates

Figure 1: Changes in the child care sector in the 2000's

Sources: Statistics Norway and regjeringen.no.

care following the reform. Second, the figure shows that the sector was severely rationed throughout the period under study.

The application process for a slot in care varies across municipalities and child care providers. A few small groups of children, such as children in foster homes and heavily disabled children, have priority for child care by law. Except from that, it is mostly up to the child care institutions themselves, and the municipalities in case of public institutions, to provide the rules for eligibility and priority in the presence of rationing. A full analysis of the allocation mechanisms in all 428 municipalities and in all the private institutions is clearly not feasible, but common priority mechanisms include birthdate, so that older children are admitted first, and sibling priority. Under these forms of rationing, the observed treatment status will generally be determined both by parental preferences and the child's ordering in the rationing mechanism, conditional on application.

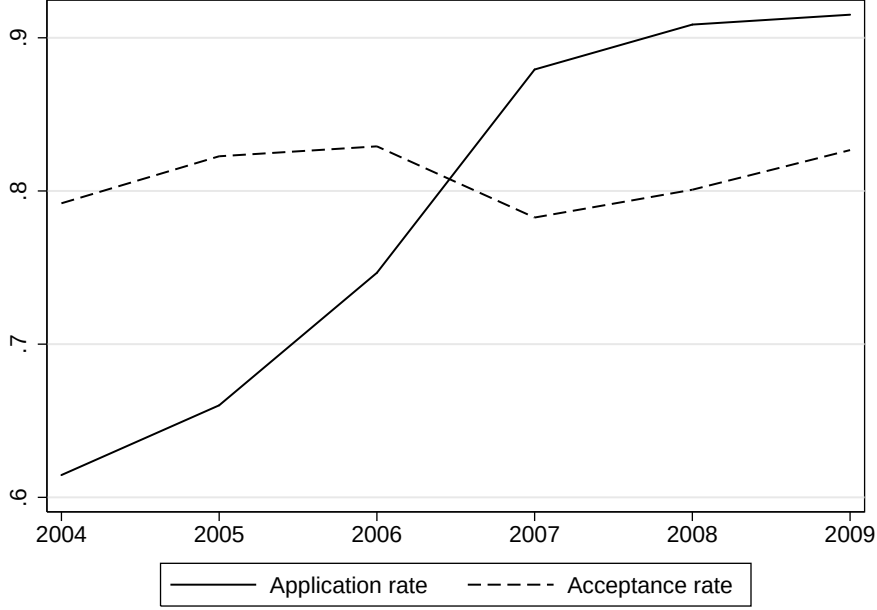


Figure 2: Rationing in the child care market

Note: Plot of municipal level application- and acceptance rates. Application rates are constructed as the number of kids in care plus the number of kids on waiting lists in care over the total population of 1–2 year olds, while the acceptance rate is the share of the applying kids that obtained a slot.

As illustrated in Figure 1d, the reform resulted in a sharp increase in municipal coverage rates for 1 and 2-year olds. Over a nine year period, coverage for toddlers increased by over 40 percentage points, from 37% in 2001 to 80% in 2010. The following empirical analysis exploits this expansion to evaluate the heterogeneous impact of early child care use.

4 Empirical strategy

The child care sector saw an explosive growth in coverage for toddlers following the child care concord. This suggests using the variation in the expansion across municipalities and over time as instruments for individual enrollment. To this end I use the coverage rates for 1–2-year olds in each municipality and year in Z_{it} to instrument for enrollment. In a rationed market, aggregate changes in coverage must be driven by changes in supply, because a mother that changes her demand for care will either not get a slot or crowd out another child. In the set of controls, I include year- and municipality-fixed effects to ensure that I only exploit the changes over time, not the differences in levels between municipalities. To further guard against observed aggregate coverage rates being driven by changes in demand that could correlate with future test scores, I control for a measure of aggregate demand obtained from the waiting list data.

Even though there is a reform that led to differential expansion in child care over time

and between municipalities, I do not claim that this expansion is exogenous. The reform itself was universal and there was no allotted budget to be allocated among applying institutions - all eligible child care institutions received the same subsidies. This removes any endogenous selection of projects by fiscal authorities, which is more likely in a setting where the institutions apply for a limited amount of funds such as in Felfe and Lalive (2015).

The obligation to expand child care to meet the demand fell on the municipalities, who expanded care directly or with the help of private suppliers. The opening of new child care centers involve complex decisions by municipalities and private suppliers, but coverage needed to be expanded in most municipalities due to large undersupply, as shown in section 5. According to a report on the progress towards the goal of full child care coverage (Asplan Viak, 2007), the most common reasons for undersupply reported by the municipalities themselves were a) demographic reasons, particularly unexpected changes in the number of children, b) local geographic mismatch of supply and demand, c) unexpected increases in demand and d) unexpected delays in construction projects. Furthermore, the annual reports on the progress towards full child care coverage (Asplan Viak, 2004-2010) show that a lot of municipalities overestimate their own progress by annually reporting to reach full coverage the following year, only to report the same one year later.

The validity of my estimation procedure relies on the orthogonality of child care expansion to underlying future trends in test scores, conditional on controls. Even though I cannot claim that the expansion is exogenous, I argue that the residual variation following the reform, given these hard to predict barriers to local expansion and conditional on an extensive set of controls, mimics an expansion that is not systematically related to trends in test scores. I support this notion using a set of robustness checks, most notably the model that includes municipality-specific linear time trends in an attempt to control for underlying trends in test scores that differ between municipalities.

In addition to fixed effects by municipality and year and the aggregate demand measure, I include in the set of controls the following individual covariates: A full set of birth month dummies, dummies for the child's immigrant status and gender, a set of fixed effects for mother's education⁶, mother's age and age squared and mother's log earnings one year before giving birth.⁷ All controls except mother's log earnings is measured in the base year, which is the year when the child turns 2. The broad immigrant dummy is equal to one if the child itself, the mother or the father has immigrated to Norway.

⁶This is based on indicators for highest registered attained education: Primary education (9 years), high school/upper secondary education (12), some college (13), bachelor (16) and master and above (18+). I also include a separate dummy for missing education rather than excluding a few percent of the sample where education is unknown.

⁷Instead of excluding the relatively few observations with missing earnings or earnings smaller than or equal to 0, I set these log earnings to 0 and include two separate dummies to account for these cases.

To give a graphical representation of this identification strategy, figure3 present the relative growth of enrollment and central controls around the year of the biggest growth in the instrument in a way that resemble an event study plot. I first remove the municipality fixed effect from each variable, then recenter the data so that year 0 is the year with the largest growth in the residual instrument. I then plot the average of residual enrollment and a few central controls around this year to get an idea of the shocks to enrollment. In all panels, the dashed line represents the average of the residual instrument, and the solid line the average of the residual enrollment or covariate. Unsurprisingly and by construction, we see large growth in the instrument between year -1 and 0, otherwise it is relatively flat. The effect of this shock to coverage is seen in enrollment (top left): Around the same time, average enrollment increased more rapidly than both before and after the shock. Reassuringly, there is no sign of a similar shock in immigrant share (top right), maternal age or education (middle) or maternal earnings and child gender (bottom). This indicates that there were not large shocks to other covariates around the same time as the expansion of care, and lends some support to the identification strategy.

In my baseline specification for the marginal treatment effects, I use a polynomial marginal treatment effects model, and verify that the results are largely confirmed when using other MTE models in section 7. When using the polynomial MTE models, specify

$$k(u) = \sum_{k=1}^K \pi_k (u^k - \frac{1}{k+1})$$

$$\Rightarrow K(p) = \int_0^p k(u) du = \sum_{k=1}^K \pi_k \frac{p(p^k - 1)}{k+1}$$

And estimate the conditional expectation of Y from (4), using the above expression for $K(p)$. I use a linear probability model to estimate the propensity score, but verify in section 7 that the results are similar when using a logit or probit model. Likewise I investigate the robustness to the baseline specification of a quadratic MTE with $K = 2$. All estimates are performed in a self-written Stata program (`mtefe`) that are available on request - see Andresen (2016) for documentation.

All standard errors are clustered at the municipality level, and are estimated by bootstrap. For each bootstrap replication, the first stage is re-estimated so that the uncertainty in the propensity score estimates are reflected in the standard errors of the MTE estimates.

5 Data and descriptive statistics

My data comes from high-quality Norwegian administrative registers that cover the entire resident population. Using unique person identifiers, I can follow people over time and link parents to children. For basic demographic variables such as date of birth, municipality

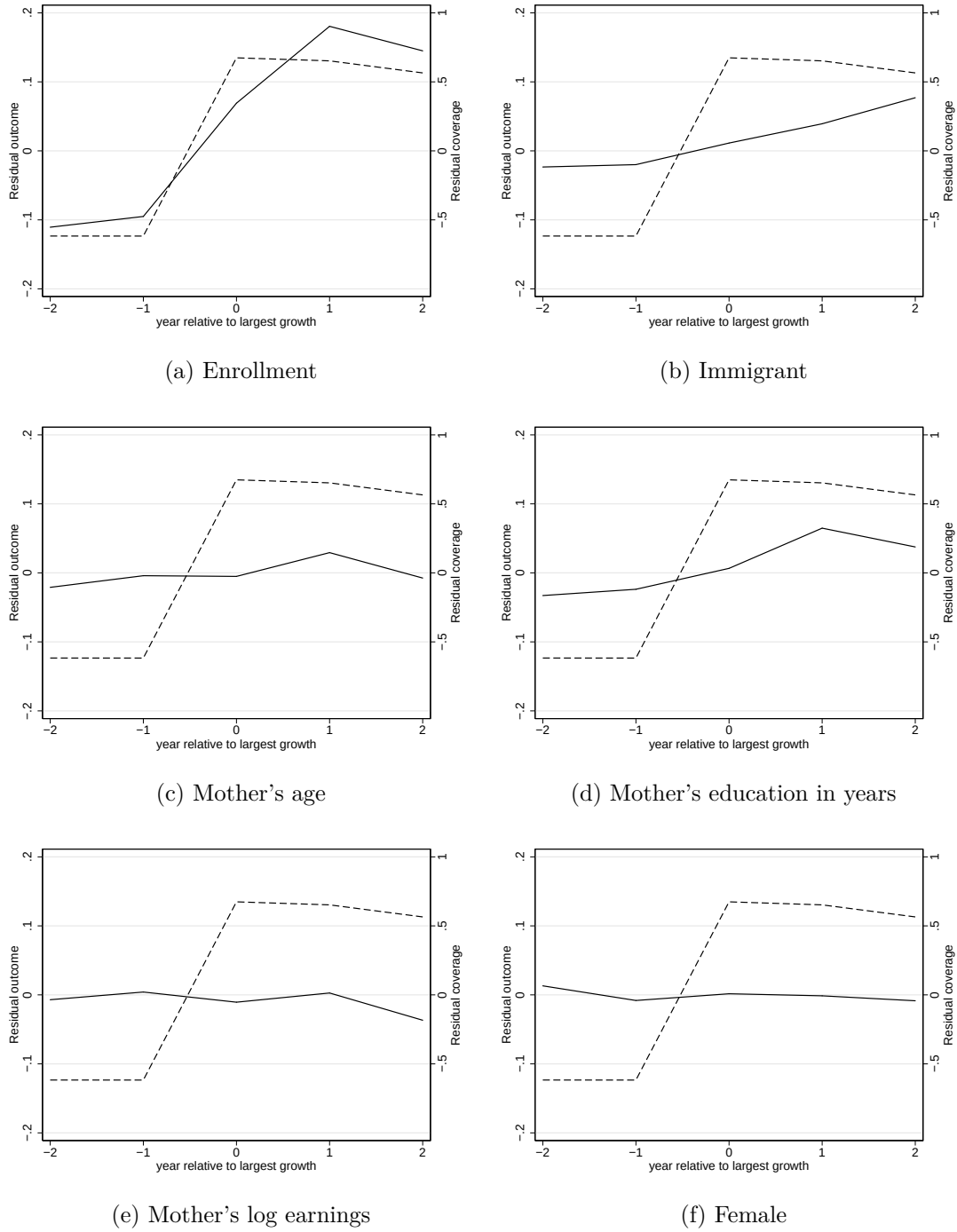


Figure 3: Event study graphs

Note: These event study graphs describe the trends in enrollment and central controls around the year of the biggest growth in the instrument. First remove municipality fixed effects from all variables, then recenter the data by municipality so that year 0 is the year of the biggest growth in the instrument. Then plot the residual instrument against the residual enrollment or control variable. The scale of the outcome variables (left axis) is 1/5th of a standard deviation of the variable in the sample. The scale of the instrument (right axis) is one standard deviation of enrollment.

of residence and family relations, I rely on data from the Central Population Register. I supplement these with income and tax information for the mother from tax records.

Test scores in math, reading and English are taken from education registers. These cover the full population, and contain test scores on standardized, nationwide tests in the three subjects taken during the autumn of the fifth school year. Tests are mandatory, and results are missing for less than 10% of the students that are enrolled to take them. More than 90% of this absence is registered as excused, probably due to sickness at the day of the test.

The most unique data material for this paper, however, is the detailed data on *child care coverage, enrollment and rationing*. The child care coverage data comes from administrative reports filed by every child care center by December 15th each year. They state the number of children that occupy a slot in their institution, separately by age. These reports form the basis of government subsidies, so that there is reason to expect them to be precise.

In addition, I utilize data from the cash for care-register to obtain information on individual child care enrollment on a monthly basis. Every child between 12 and 35 months of age that does not occupy a subsidized slot in care is entitled to a relatively substantial cash for care benefit.⁸ As long as all children who do not use care take up this benefit, and knowing that you are eligible for the benefit the first month that you start care, I know precisely which children are in care. It is also possible to take up part cash for care if your child attends part-time care. Based on the rates from the registers, I construct measures of full-time-equivalent months of care for each month from age 12 to 35 months. This measure of child care use contains information on both the extensive and intensive margin of treatment.

For the marginal treatment effects framework I need a binary treatment indicator. Based on the continuous treatment measure, I can construct a starting age measure: the first month that I observe at least some child care use. Given that child care is largely an absorbing state, there is little need to worry that an important share of kids go back and forth between modes if care.

The distribution of the calendar month of child care start is plotted in figure 4. The first panel shows the share of kids starting sometime in the year after their birth year, the second for the kids starting in the calendar year two years after birth and so on. Note that for approximately 15% of the sample, I cannot exactly determine the start month. This is because I do not observe any child care use up to 35 months of age, and after that the child is no longer eligible for the benefit. I therefore know that the child did not start before it's third birthday, but not exactly when he or she started.⁹ This worry does not

⁸The benefit varied between NOK 3,000 and NOK 3,657 per month in the period, or 3,880 to 4,716 in May 2016 NOK. This is equivalent to around 470\$ to 570\$ using a May 2016 exchange rate.

⁹In principle, there could be a similar worry for very early start dates as I do not observe child care use before 12 months of age. However, parental leave benefits in Norway extend 49 weeks, making it highly

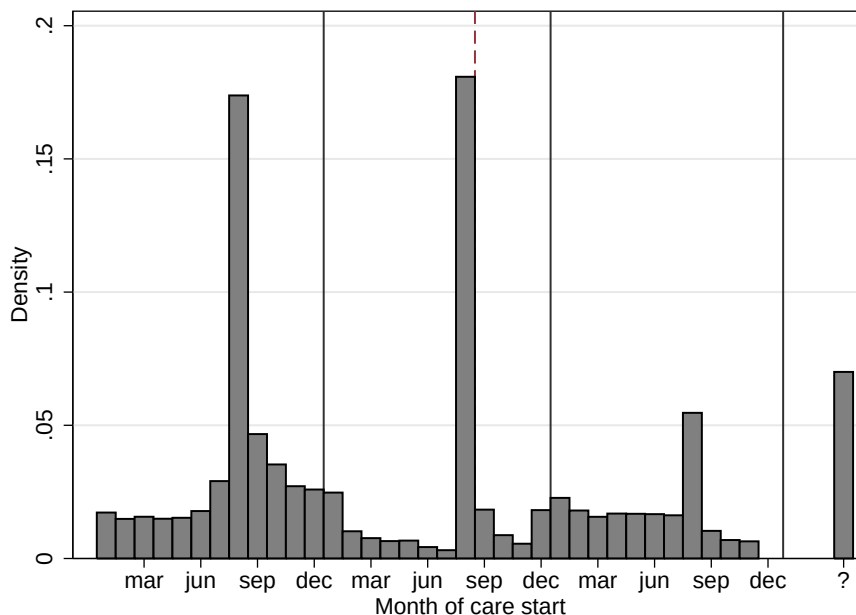


Figure 4: Distribution of calendar month of child care start

Source: Cash for care register. Note that we cannot observe exact start for children who start after their third birthday, as these are no longer eligible. These are grouped in the “?” bar.

affect the classification of the treatment indicator.

Figure 4 unsurprisingly documents clear seasonal patterns in child care enrollment: A lot of children start in August each year simply because slots in care are freed up as older children advance to school. The probability of securing a slot in care at this time also very likely depends on time of birth, as this is a common priority mechanism. If timing of birth affect school performance, it is crucial to control for this seasonal pattern in enrollment.

Having seen this, I define early care as having started care in or before August in the second calendar year after birth. Given this definition, the possible errors in the classification of children starting before their first or after their third birthday is of little importance. The kids considered as treated according to this definition are the ones to the left of the dotted line in figure 4. A conservative estimate of the difference in the starting ages indicate that the treatment group is exposed to at least 16 more months of care than the control group according to this definition.

Lastly, I have unique data on the degree of rationing in the municipalities. As a way to monitor the progress towards the goal of full child care coverage, the Ministry of Education surveyed each municipality in the years 2004 to 2009, asking them about the number of children by age on the waiting list for a child care slot. This allows me to assess the degree of undersupply of care in each municipality and year.

unlikely that a considerable share of children start much before 12 months. If there are such children, I would wrongly classify them as starting the month they turn 1 year old.

I use these data to form two rates: the *application rate* and the *acceptance rate*. The application rate is a measure of the share of children aged 1 and 2 that applies for child care, while the acceptance rate is the share of the applying children that occupy a slot in care. Until 2006, the waiting list data only separates children based on whether they are below or above 3 years of age. Since parental leave benefits extend almost a full year, however, I assume that there are no 0-year olds on the waiting lists, so that I can treat the waiting lists for below 3 as consisting only of 1- and 2-year olds.¹⁰

The average application- and acceptance rates using these definitions are plotted in figure 2. We see primarily a large increase in the application rates that reflect increased demand for care. Despite this, the acceptance rate has not declined, as evident by the relatively flat acceptance rate over time, because of the large increase in supply. In total, a little less than 20% of the children that applied were not accepted, and thus were rationed. When looking at the municipalities, 70% of the municipalities had more than 5% of their 1- and 2- year olds on waitings lists for care, indicating considerable rationing in the majority of municipalities.

Sample

I begin with all 227,000 2-year olds alive and resident in Norway in the years 2002-2007. I drop a handful of children with missing information for the mother. I also drop 9,500 children who were not enrolled in fifth grade in the correct year, when outcomes are measured. This is mostly due to early or late school start, which is relatively uncommon, but at the discretion of the parents.

The remaining 217,500 children were scheduled to take the tests in the years 2012-2015. Of these, around 17,000 were absent from at least one of the tests, and are dropped to keep a consistent sample. About 90% of this absence is registered as excused. Lastly, I have to drop a few handful of observations from very small municipalities due to collinearity. This leaves a final sample of around 200,000 children.

Descriptive statistics

Table 1 provides descriptive statistics. The sample is balanced with respect to gender, and around 68% of the sample is treated in the sense of starting child care in or before August the year they turn two. During their second and third year of life, the sample children spend on average around 10.7 full-time-equivalent months in care. We also see fair variation in some key covariates on the mother, particularly years of education, log earnings the year before birth and age. Coverage rates for 1–2-year olds average 58% in the period of study, but varies considerably. In contrast, the measure application rates is

¹⁰On average, only 3 per cent of the 0-year olds occupy a slot in care in my sample, although in principle more could be applying, but rationed.

Table 1: Descriptive statistics

		mean	sd	min	max
Child	Female	0.50	0.50	0	1
	Immigrant	0.086	0.28	0	1
Early care	Treatment dummy	0.68	0.47	0	1
	Months of care 13-35	10.7	8.62	0	23
Mother	Education (years)	13.6	2.98	6	21
	Log earnings, $t - 3$	12.0	0.97	1.39	15.0
	Age	32.2	5.02	16	56
Aggregate care	Coverage CC_{kt}^{12}	0.58	0.13	0.061	1
	Application rate	0.73	0.17	0.17	1
Test scores	Reading	20.0	6.40	0	33
	Math	25.3	9.23	0	45
	English	27.1	10.6	0	50
N		200,781			

Note: Descriptive statistics for the main sample. Sources: Population, tax, cash for care, education and child care registers.

on average 73%, indicating severe rationing on average.

My outcome variables are test scores in fifth grade. Non-standardized averages are shown in table 1. These are graded on scales that vary across subject and year, a few times also between exam sets within subject-year. Figure 5 provides kernel density estimates of the three outcome variables, fitted with a normal distribution. Given that the fit of the normal distribution is relatively good, I will work with normalized test scores in the following to ease interpretation, so that all results can be interpreted as standard deviations.¹¹

6 Results and analyses

OLS and IV results

I start by presenting the OLS and 2SLS estimates in table 2. Clearly, there are strong selections issues with the OLS estimates that excludes any causal interpretation of the estimates, but taken at face value, there seem to be modest average gains from early care. Test scores are .08 higher for reading, .07 higher for math and .04 higher for English, all

¹¹Although it does not seem to matter, I standardize within year, subject and problem set to account for any differences between problem sets.

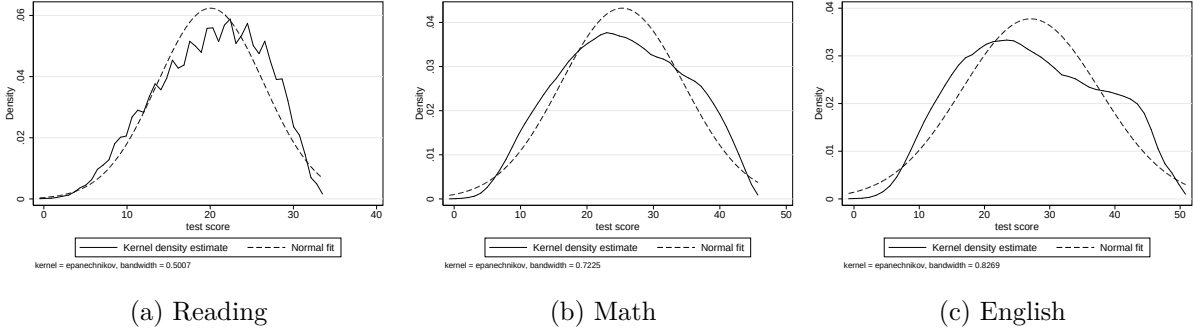


Figure 5: Kernel density plots of outcome variables

Note: Kernel density estimates of the distribution of test scores using an epanechnikov kernel.

Rule-of-thumb bandwidth.

significant at the 1% level.

If the IV assumptions hold, the IV strategy mitigates the selection problem in the OLS estimates. In the second panel of table 2, we find the first stage estimates. Both coverage rates have strong and significant positive impacts on the individual enrollment, as evident by the .18 and .27 coefficients. The F -statistic is above 70, indicating a strong and relevant first stage. Effects are similar across the two samples.

Moving to the last panel of table 2, we find the second stage IV estimates. These indicate *negative* results of early care, even significant at the 10% level for reading. Taken at face value, early care leads to reductions of .3 standard deviations in reading, .04 standard deviations in math and .08 standard deviations in English, but note that this is only significant for the reading estimate. This indicates that the OLS is troubled by positive selection: Children who would have higher test scores independently of child care use, are more likely to enroll.

These estimates are large and alarming from a policy perspective, but we need to remember that these are local average treatment effects: Average treatment effects among the people induced to start child care because of the increased availability. As we shall see, the compliers to these instruments are quite different from the average child, generating negative LATE estimates, and taking these as evidence of no or negative effects of child care on test scores in general is potentially very misleading.

Common support and first stage results

The full first stage results are presented in table 3. The instrument has clear predictive power for enrollment into early care. In addition, we can note some additional features of the selection into treatment. Immigrant children are around 3 percentage points less likely to enroll, while boys are around 1 percentage point more likely to enroll. Mothers' education has a positive and relatively linear effect, as evident by the coefficients on the dummies for education, with children of highly educated mothers more likely to enroll than

Table 2: OLS and IV estimates of the impact of early care on test scores

	Reading	Math	English
A: OLS: Test scores			
Early care	0.0946*** (0.0176)	0.0778*** (0.0130)	0.0415*** (0.00851)
B: First stage IV: Early care use			
Coverage	0.510*** (0.0503)		
F	102.8		
C: Second stage IV: Test scores			
Early care	-0.109 (0.127)	-0.0218 (0.130)	-0.0894 (0.138)
Individual controls	✓	✓	✓
Municipality FE	✓	✓	✓
Year FE	✓	✓	✓
Birth month FE	✓	✓	✓
N	200,781		

Note: Panel A shows the OLS estimates of test scores on an indicator for early care and controls. Panel B shows the first stage IV estimates of enrollment on controls, and panel C shows the 2SLS estimate of early care on test scores. Standard errors in parentheses, clustered at the municipality level. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

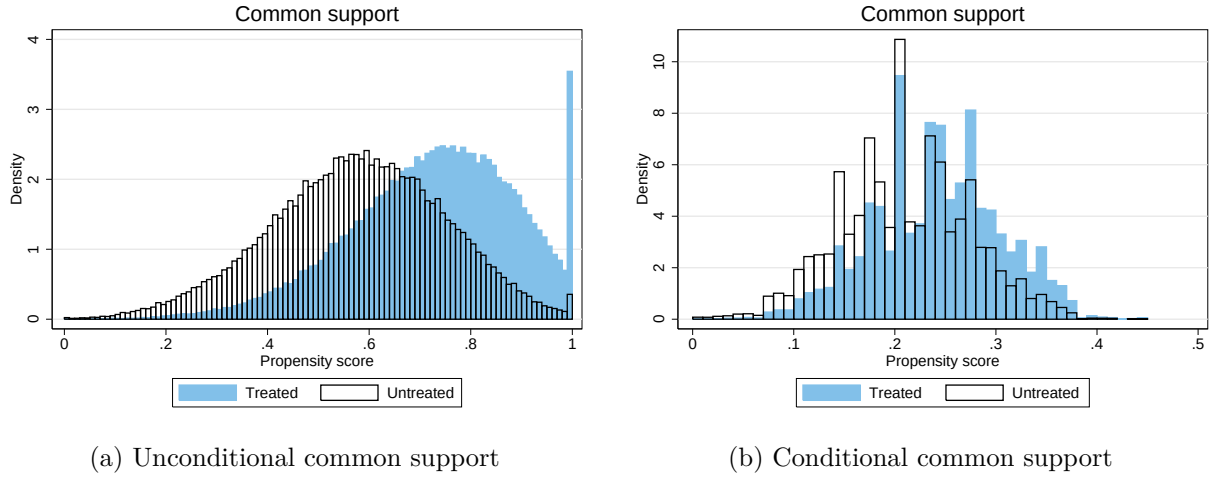


Figure 6: Common support plot

Note: Plots of the common support in the treated and untreated samples. Rightmost plot illustrates the variation driven by the instruments by predicting propensity scores from the first stage after setting all continuous regressors to their mean, all fixed effects to the base levels and demeaning the instrument.

children of mothers with low education. Mother's age does not seem to drive enrollment, and neither does the aggregate demand measure. Children of mothers with higher earnings seem to be more likely to enroll.

Figure 6 plots the resulting propensity scores for the treated and untreated samples. The linear probability model generates propensity scores above 1 that are manually adjusted to 1, and, to a lesser extent, propensity scores below 0. I show in section 7 that using the probit or logit model in the first stage yields equivalent MTE estimates

The left panel of figure 6 shows the unconditional common support. This variation includes all the variation in $P(Z)$, including the variation induced by X . We can see how the treated individuals have a distribution of propensity scores that lies higher than the untreated children. In contrast, the right panel illustrates the variation in common support driven exclusively by the variation in the instrument. To plot this, I predict the propensity scores from the first stage after first setting all continuous covariates to their mean, all fixed effects to their base levels and demeaning the instrument. This leaves only the variation in the instrument to generate variation in propensity scores. Obviously, this variation is much smaller than the unconditional common support, and illustrates why relaxing the additive separability assumption is often unfeasible in applied work. Nonetheless, we see fair variation in the propensity scores driven by the instrument.

Lastly, note how there is limited support at the left tail of the propensity score distribution. This is not an issue in the baseline parametric model, but when using semiparametric methods, it is necessary to trim the common support and estimate the MTE only at the points where we have considerable support in both samples.

Table 3: First stage estimates

Parameter		Coefficient	
Mother's age		-0.00471	(0.00403)
Mother's age ²		0.0000372	(0.0000580)
Immigrant		-0.0296***	(0.00436)
Female		-0.00983***	(0.00217)
Mother's log earnings $t - 3$		0.0177***	(0.00106)
Mother's earnings missing		0.176***	(0.00820)
Mother's earnings ≤ 0		0.0719***	(0.0266)
Mother's education	Upper secondary	0.0868***	(0.00921)
	Some college	0.158***	(0.0101)
	Bachelor	0.215***	(0.00988)
	Master or higher	0.290***	(0.00973)
	None/unknown	-0.0606***	(0.0145)
Coverage rate		0.510***	(0.0503)
Application rate		-0.00625	(0.0326)
Constant		0.161*	(0.0873)
Municipality FE		✓	
Year FE		✓	
Birth month FE		✓	
N		200,781	

Note: Table shows coefficients from a linear probability model of enrollment on Z . Standard errors in parentheses, clustered at the municipality level. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

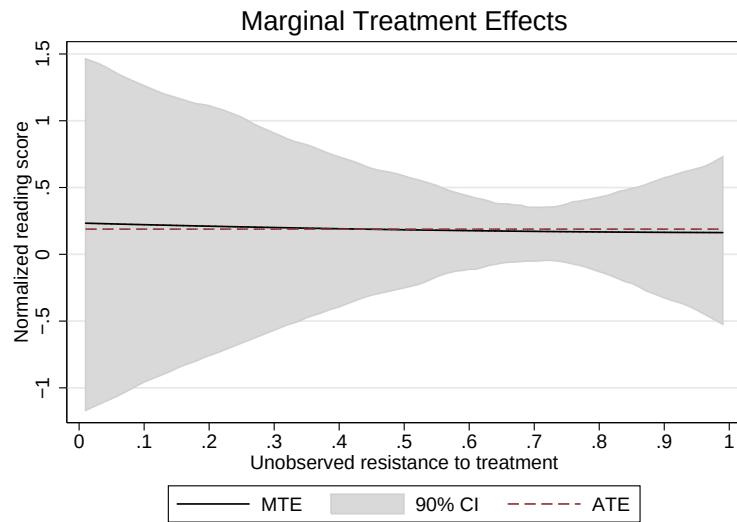
MTE estimates

The estimated marginal treatment effects, plotted at the mean of X in the sample, is depicted in figure 7. For reading, the MTE curve is almost flat, and the average treatment effect is around .19 standard deviations, although this is not significantly different from zero. A joint test of the coefficients on p cannot reject that it is flat. Consequently, there doesn't seem to be heterogeneity along the unobserved distaste for treatment. More or less the same conclusion holds for English: Although the MTE is downward sloping, we cannot reject that the coefficients on the nonlinear terms are 0, neither separately or using a joint test. The average treatment effect for English is around the same as for math, .2 SD, but not significantly different from zero.

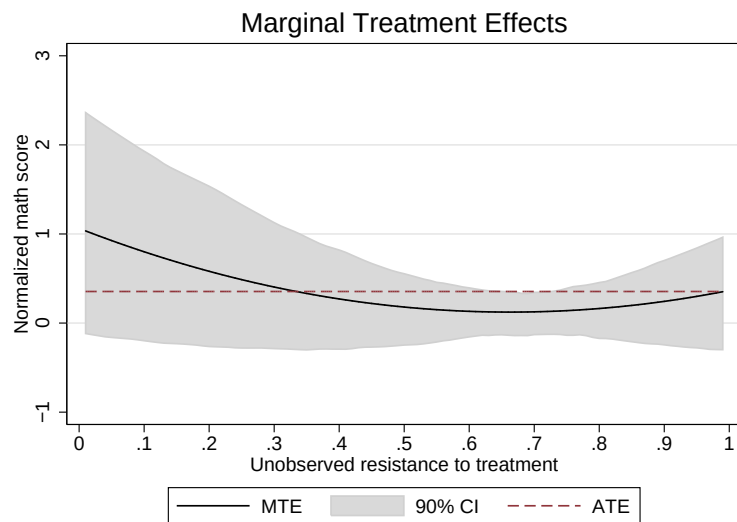
For math, however, the story is different: The marginal treatment effects are downward sloping. Both coefficients in $k(u)$ are significant at conventional levels ($p = 0.04$ and $p = 0.02$, respectively), and a joint test of their significance is significant ($p = 0.05$). The treatment effects are large and approaching significance at the beginning of the U_D distribution, indicating treatment effects close to one full standard deviation for these children. As we move up the distribution of distaste for treatment, marginal treatment effects drop and approach zero, then increase somewhat for the kids with the largest resistance.

Downward sloping MTEs is consistent with an economic model where there is selection on gains: Children who have a large and positive unobserved gains component, are more likely to select into treatment precisely because their gain from doing so is large. Whoever makes the decision to enroll the child, be it the parents or the child care centers, could be making that decision partly based on knowledge of the unobserved gains of the children. This pattern of the MTE is in line with Felfe and Lalive (2015), but in contrast to Cornelissen et al. (2016) and Kline and Walters (2015).

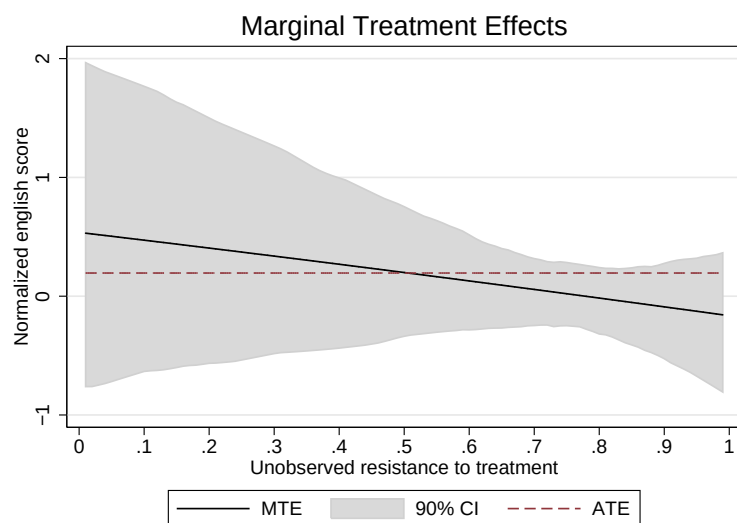
Some authors (Felfe and Lalive, 2015; Cornelissen et al., 2016) use the restriction that the municipality-, year- and/or birth month fixed effects are the same across treatment states. Although these authors also employ average quality indicators as controls, which could alleviate some of the worry that this restriction is arbitrary, I employ a more flexible specification where the fixed effects can differ in the treated and untreated state. This allows, for example, that some municipalities may produce very poor outcomes in the untreated state, perhaps by having other child and family policies that does not encourage child development, while still producing good outcomes in the treated state, perhaps by having high quality child care institutions. To evaluate whether this restriction matter in my case, I employ the restriction that a) the municipality fixed effects are the same across treatment states or b) that the municipality-, year- and birth month fixed effects are the same across treatment states. The resulting MTEs are plotted in figure 8 together with the baseline MTEs. Clearly, the restriction matters: The MTEs using these restrictions



(a) Reading



(b) Math



(c) English

Figure 7: Marginal treatment effects of early care on normalized test scores at age 10

Note: Marginal treatment effects at the mean of X .

Table 4: Parameters in the MTE model

Parameter		Reading		Math		English	
		$\hat{\beta}_0$					
Mother's age		-0.0534***	(0.0127)	-0.0388***	(0.0151)	0.000819	(0.0150)
Mother's age ²		0.00105***	(0.000192)	0.000781***	(0.000241)	0.000170	(0.000235)
Immigrant		0.0566*	(0.0324)	0.0494	(0.0381)	0.0767**	(0.0369)
Female		0.103***	(0.0178)	-0.158***	(0.0172)	0.0272	(0.0225)
Mother's log earnings		0.00486	(0.00749)	-0.00746	(0.00793)	-0.00502	(0.00844)
Mother's earnings missing		0.0611	(0.108)	-0.0125	(0.116)	0.0430	(0.0897)
Mother's earnings ≤ 0		0.0572	(0.202)	-0.322	(0.201)	-0.201	(0.165)
Application rate		-0.0411	(0.174)	-0.0921	(0.180)	-0.180	(0.166)
Mother's education	Upper secondary	0.216***	(0.0439)	0.263***	(0.0465)	0.159***	(0.0483)
	Some college	0.324**	(0.126)	0.451***	(0.113)	0.251**	(0.118)
	Bachelor	0.513***	(0.102)	0.574***	(0.0966)	0.400***	(0.109)
	Master or higher	0.727***	(0.128)	0.752***	(0.137)	0.657***	(0.155)
	None/unknown	-0.0919*	(0.0484)	0.00673	(0.0531)	0.115*	(0.0591)
Constant		-0.0237	(0.289)	-0.193	(0.297)	-0.680**	(0.322)
		$(\widehat{\beta_1 - \beta_0})$					
Mother's age		0.0681***	(0.0230)	0.0901***	(0.0242)	0.0145	(0.0215)
Mother's age ²		-0.00129***	(0.000334)	-0.00163***	(0.000361)	-0.000236	(0.000330)
Immigrant		-0.0742	(0.0493)	-0.0869	(0.0563)	0.227***	(0.0489)
Female		0.0182	(0.0263)	0.0379	(0.0235)	-0.0613*	(0.0330)
Mother's log earnings		0.0133	(0.0107)	0.0414***	(0.0111)	-0.000330	(0.0108)
Mother's earnings missing		-0.0457	(0.144)	0.164	(0.152)	-0.0144	(0.133)
Mother's earnings ≤ 0		0.209	(0.378)	0.839**	(0.383)	0.421	(0.316)
Application rate		-0.0462	(0.241)	0.0749	(0.253)	0.219	(0.228)
Mother's education	Upper secondary	-0.0574	(0.0641)	-0.0956	(0.0710)	-0.107	(0.0698)
	Some college	-0.0570	(0.174)	-0.222	(0.162)	-0.128	(0.160)
	Bachelor	-0.0464	(0.140)	-0.0979	(0.131)	-0.156	(0.139)
	Master or higher	-0.0228	(0.174)	-0.0339	(0.189)	-0.198	(0.188)
	None/unknown	0.133	(0.103)	0.0449	(0.0937)	-0.0347	(0.101)
Constant		-0.381	(0.781)	-0.202	(0.733)	0.543	(0.698)
		$\hat{\pi}$					
π^1		-0.130	(1.521)	-2.834**	(1.402)	-0.650	(1.485)
π^2		0.0581	(0.987)	2.138**	(0.873)	-0.0518	(1.035)
Birth month FE		✓		✓		✓	
Municipality FE		✓		✓		✓	
Year FE		✓		✓		✓	
N				200,781			

Note: $\hat{\beta}_0$, $(\widehat{\beta_1 - \beta_0})$ and $\hat{\pi}$ from parametric polynomial MTE model. Bootstrapped standard errors clustered at the municipality level. Estimated using Local IV. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

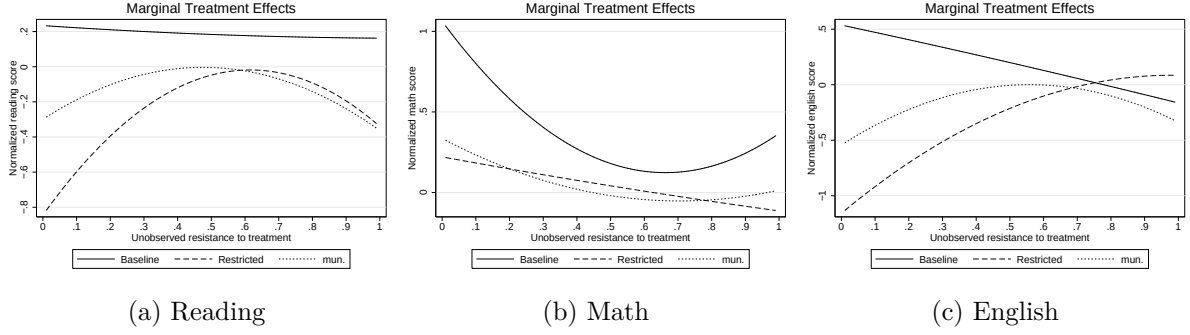


Figure 8: Restricting the fixed effects

are drastically different from the baseline MTEs found with the flexible specification. Based on this, restricting the fixed effects to be the same across treatment states drives the results in my data.

All parameters in the baseline MTE model is reported in table 4. The β_0 coefficients can be interpreted as effects of covariates in the absence of treatment. Unsurprisingly, the coefficients in the first panel document a clear positive relationship between mother's education and school performance: Children of better educated mothers do better than less advantaged peers, particularly for mothers with university education. Surprisingly, immigrant children do better majority kids in absence of treatment, with approximately .05 to 0.08 standard deviations better test scores. Girls do .1 standard deviation better than boys in reading, while boys do .16 standard deviation better in math. There is no difference between the genders in English.

Next, I am interested in how the treatment effect differ along observables, estimated as the $\beta_1 - \beta_0$ coefficients. Given the assumptions of separability between X and $K(p)$, different observables will lead to horizontal shifts in the MTE curve. The size of the shifts can be read directly out of the $\beta_1 - \beta_0$ coefficients: For example, the estimate in panel 2, row 1 indicates that immigrant children have .07 standard deviation lower treatment effect than majority children for reading, so majority catch up to the immigrant kids' better reading scores when enrolling early in care. This also hold for math, with .09 standard deviations lower treatment effect, but is reversed for English, where immigrants have .23 higher treatment effects. This might seem puzzling and in contrast to current policy proposals aiming at increasing immigrant children's enrollment, where it is claimed that enrollment equalizes differences between majority and minority kids. According to these results, it does, but it is actually the majority who's catching up to the immigrants in reading and math, rather than the opposite. For English, the two groups do approximately the same in the absence of early care, while immigrants do much better when enrolled. Remember, however, that these are partial effects: All other variables are held constant at the mean in the full sample.¹²

¹²Another way to measure the full difference between the treatment effects in the majority and minority

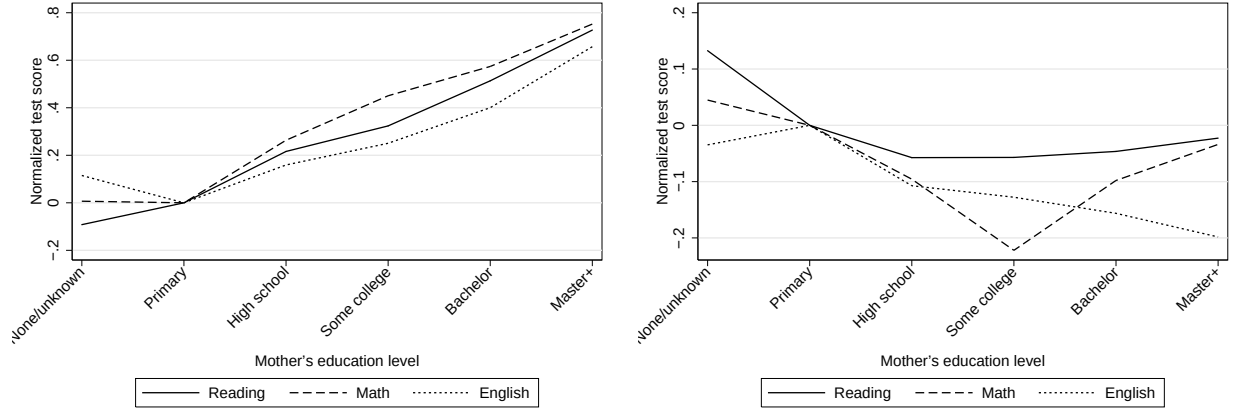
Girls seem to have .06 SD lower treatment effect in English, but not in math and reading. This is an indication that while girls do better in the absence of treatment in English, boys catch up when enrolled in care. Boys' advantage in math or girls' in reading in the untreated case, however, is not equalized the same way by early care. Maternal age does not seem to matter for English, but for math and reading, children of relatively young mothers (26-27 years) have the most to gain from early care, while children of older and younger mothers gain less.

Perhaps most interesting among the observable characteristics are the dummies for mother's education. Mother's education is often used as a proxy for socioeconomic status, and is an important dimension along which to evaluate the claim that child care equalizes differences in early endowments. Along with a separate dummy for missing education, the education dummies indicate the relative size of the treatment effect for children of mothers of different education levels, relative to the baseline level of only 9 years of primary schooling. Table 4 shows that none of these dummies are significant at any conventional level. Figure 10 shows differential effect graphically. The gradient seems to be downward sloping for English, relatively flat for reading except for the increased treatment effect for children of mothers with unknown education levels, and nonlinear for math, with the smallest treatment effect for mothers with some college. Thus, there is some support for the idea that child care benefits disadvantaged children more than advantaged ones for English, but mixed evidence for the other two topics. I cannot reject a joint test of significance for the education dummies for either of the outcomes.

To further address this idea, I estimate the MTE using the separate approach as suggested by Heckman et al. (2006), outlined in section 2. Asymptotically, this estimation method should yield comparable estimates as the Local IV method, but in smaller samples they might deviate.¹³ Using the separate approach, I estimate all the parameters of the two outcome functions, not just their difference, and this allows me to plot the potential outcomes separately. These plots are found in figure 10, with the baseline specification using a second order polynomial on top and fully semiparametric estimates on the bottom. Remember that the common support is very sparse at the left tail of the U_D -distribution, so that semiparametric estimates should be interpreted cautiously up to about .2. Nonetheless, these results are quite different than the baseline estimates using Local IV for both reading and math. The MTE curves are relatively flat, maybe even upward sloping, in contrast to the Local IV results. I have no good explanation for this

population would be to investigate the total difference in the index of X : $(\widehat{\beta_1 - \beta_0})(\bar{x}_I - \bar{x}_M)$, where $\bar{x}_I$ is the vector of the mean of X in the immigrant population and $\bar{x}_M$ the same in the majority population. This will give a more relevant estimate of the effect of sorting on immigrant status when allocating care. When doing this, I find that immigrants have X -values that give them a 0.02 lower treatment effect in reading, .11 lower treatment effect in math, but a large .28 higher treatment effect in English. Therefore, this does not substantially change the conclusions from above.

¹³Preliminary Monte Carlo analysis suggest that the samples need to be relatively large before the two approaches converge.



(a) β_0 : Mother's education in the untreated state

(b) $\beta_1 - \beta_0$: Difference in treatment effects

Figure 9: Heterogeneity in mother's education

Note: Figure plots coefficients on the dummies for mothers' education.

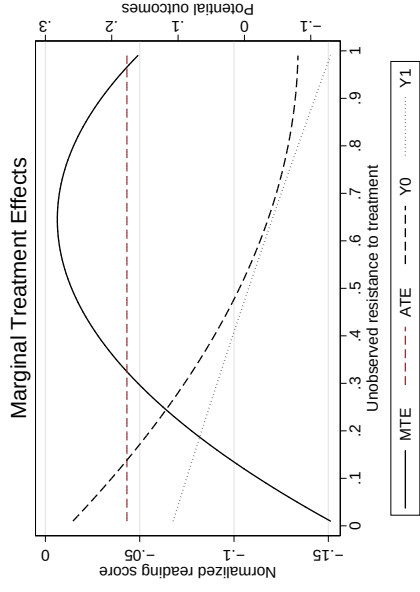
surprising finding.

Treatment effect parameters

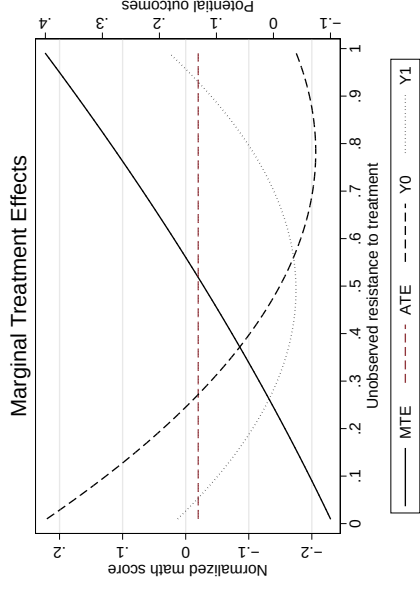
Having estimated the marginal treatment effects, we are now in a position to investigate the local average treatment effects from section 6 in detail. Having seen the main results, we know that the average treatment effect (ATE) is simply the average over the MTE curve. For math, it amounts to around .36 standard deviations increase in math scores from attending early care, significant at the 10% level. This is in stark contrast to the 2SLS estimate of -.02. The same results for reading and English shows insignificant ATEs of .19 and .2, in contrast to 2SLS estimates of -.11 and -.09.

Table 5 shows the estimated treatment effect parameters. These have been calculated as weighted averages over MTE curve for the relevant population, using the estimated weights. The top panel of figure 11 plots the MTE at the mean that we already saw together with the MTE for the compliers along the left y -axis. The compliers to the instruments are different than the main sample, and so they have a different average X . As it turns out, the average of the X among the compliers is such that the MTE is shifted considerably downwards compared to the MTE at the mean. In fact, the index $x(\widehat{\beta_1} - \widehat{\beta_0})$ is around .2 standard deviation lower for the compliers than the mean in the population for all three outcomes. Thus, the compliers are people with X -values that give them smaller treatment effects than the average. This is the first reason why the LATE differs from the ATE.

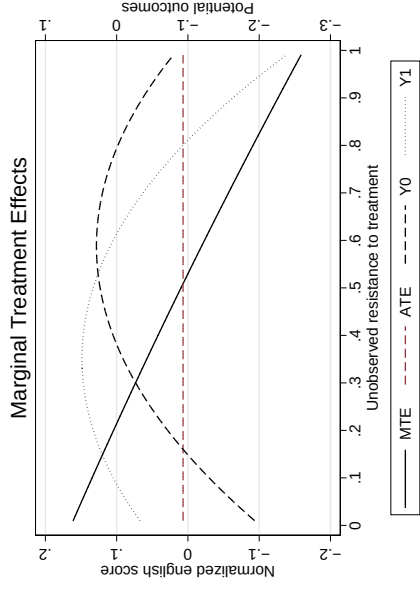
Along the right y -axis of figure 11, I plot the estimated weights that IV puts on the different parts of the U_D -distribution. In the population, U_D is uniformly distributed, but this does not hold in subsamples. As is evident from the figure, people with U_D in the middle of the distribution, from approximately .4 to .8, is much more common in the



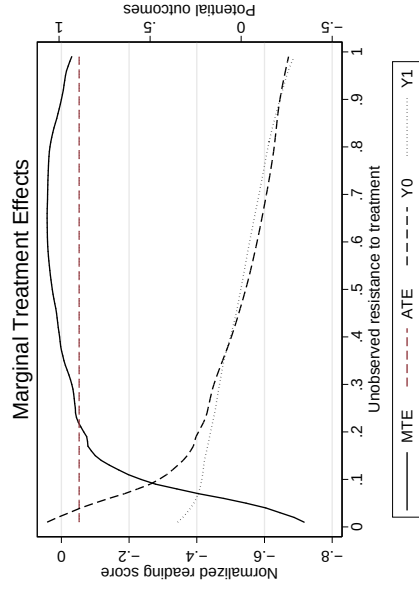
(a) Reading



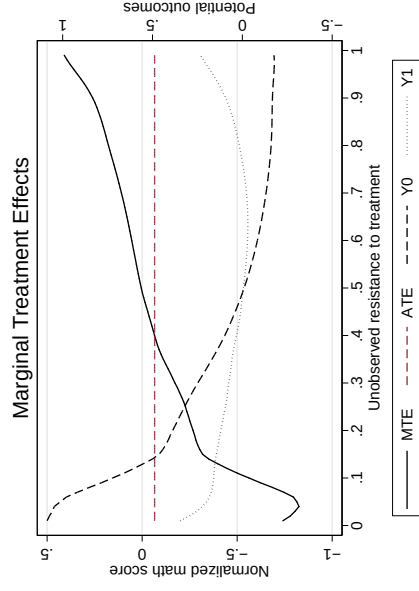
(b) Math



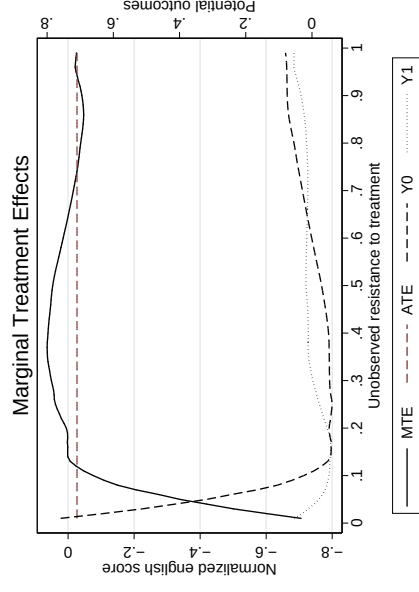
(c) English



(d) Reading, semiparametric



(e) Math, semiparametric



(f) English

Figure 10: MTEs estimated using the separate approach

complier population than in the full population. Therefore, the local average treatment effect puts a lot of weights on the MTEs of people with U_D in this interval. Because the MTE curves are downward sloping, these people have MTEs close to or even below 0, so that the small and negative part of the MTE distribution will be given a high weight when calculating the LATE. This explains why the local average treatment effect can be small and negative even when the ATE is relatively large and positive.

Furthermore, if the separability assumption holds, the LATE estimate calculated as a weighted average over the MTE for the compliers should be identical to the standard 2SLS estimate. In the figure, the 2SLS estimate is indicated with a dotted line. Compared to the 2SLS estimates in table 2, these are very close. I take this as some support for the separability assumption.

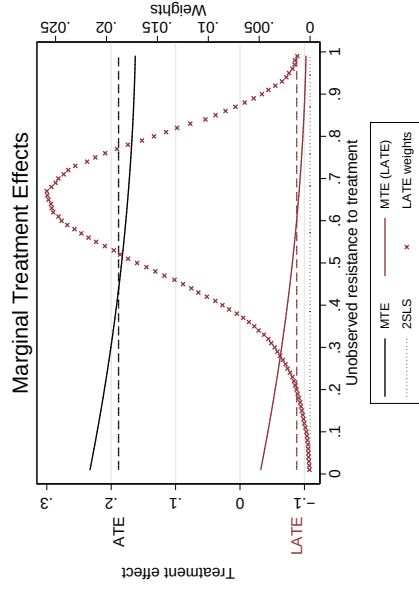
In a similar fashion, we can calculate the weights that other treatment effect parameters put on the different parts of the U_D distribution. The bottom panel of figure 11 shows the MTEs and the estimated weights for the average treatment effect on the treated (ATT) and the average treatment effect on the untreated (ATUT). Notice first that the MTE for the sample of treated children lies above the MTE at the mean for reading and math, while the MTE for the untreated sample lies below. This is another indication of selection on observable gains: The children who choose treatment have X -values that makes their treatment effect larger, and opposite for the untreated children. For English, the three MTE curves practically overlap, indicating no important selection on observable gains.

Furthermore, the weights for both parameters are plotted on the right y -axis. The ATT puts a lot of weight on the left part of the U_D - distribution. These are children with low distaste for treatment, so all else equal, they are more likely to be in the treated sample. Conversely, the ATUT puts most of the weight on the right side of the U_D distribution precisely because these children have high distaste for care: All else the same, they are less likely to be treated.

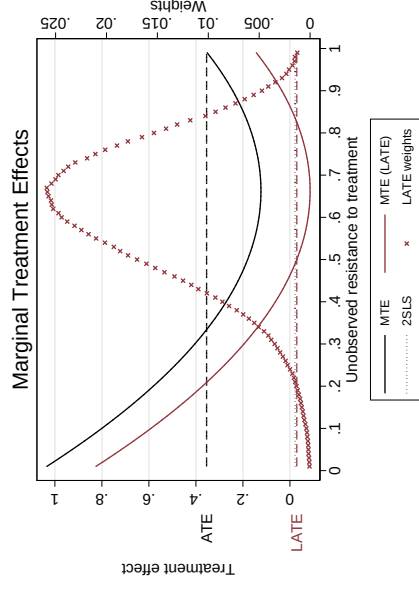
Calculating the ATT and ATUT as the weighted average over the appropriate MTE curve using these weights, we find an average treatment effect on the treated at .44 standard deviations increase in math, .2 in reading and .29 in English - large gains, although not significantly different from zero. In contrast, the ATUT estimate is .17 for both reading and math and practically zero for English: We expect there to be much smaller effects by assigning treatment to all the children not currently enrolling in care.

Using policy relevant treatment effects, I can use the same method to compute the expected treatment effect for the compliers to any imagined policy change. This assumes policy invariance (Heckman and Vytlacil, 2007) in the distribution of covariates and is based on out-of-sample extrapolation, but is a way to use the MTE model to make predictions about the effects of hypothetical policy changes.

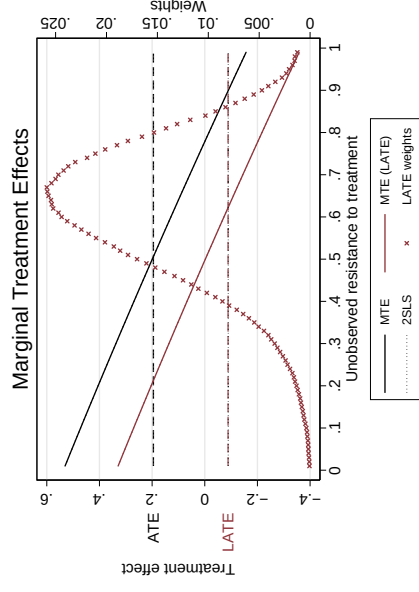
Remember that the expansion I exploit in this paper did not end in 2007, when my



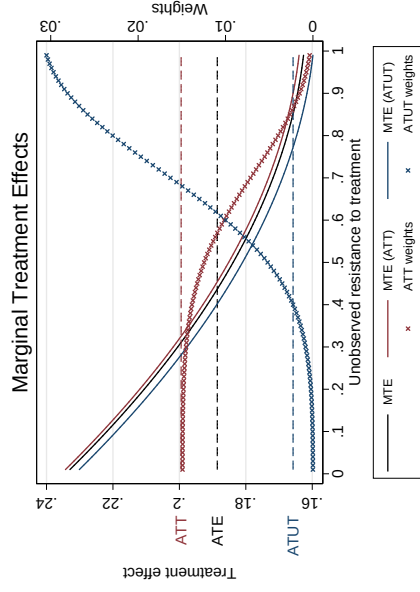
(a) LATE weights, reading



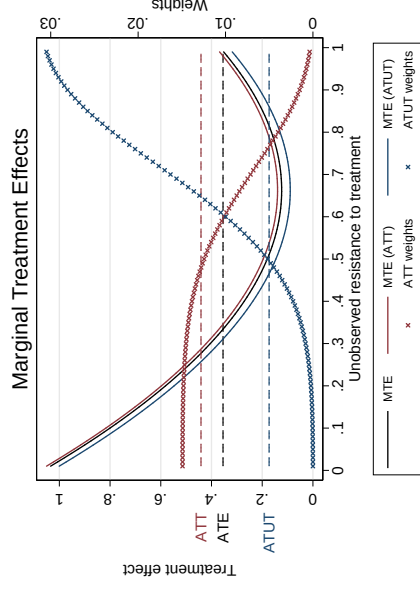
(b) LATE weights, math



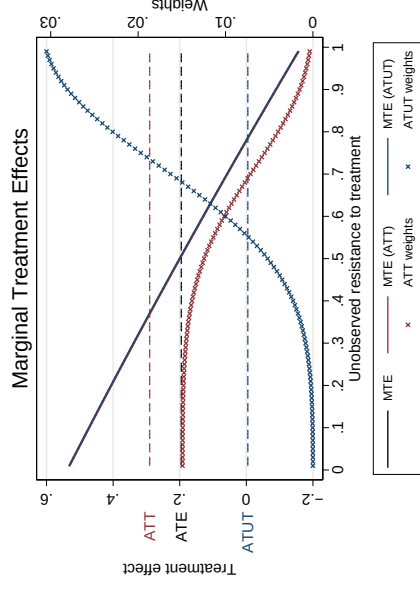
(c) LATE weights, English



(d) TT and TUT weights, reading



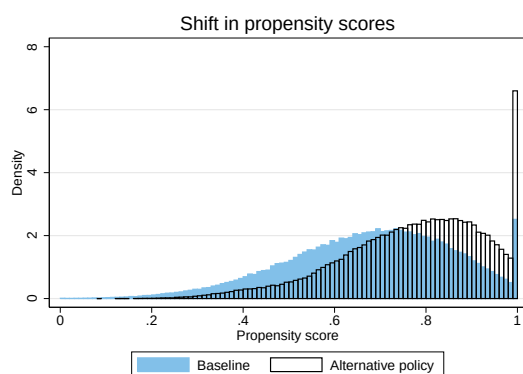
(e) TT and TUT weights, math



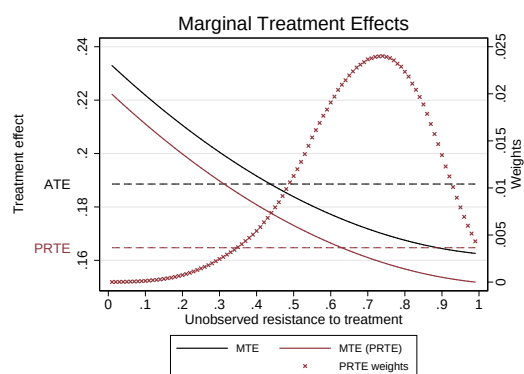
(f) TT and TUT weights, English

Figure 11: Treatment parameter weights

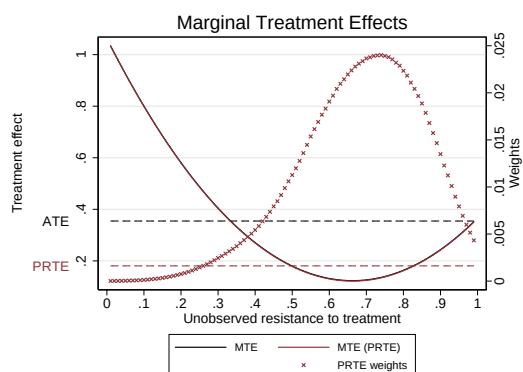
Note: Marginal treatment effects for the relevant subpopulation based on several common treatment effect parameters, together with associated weights.



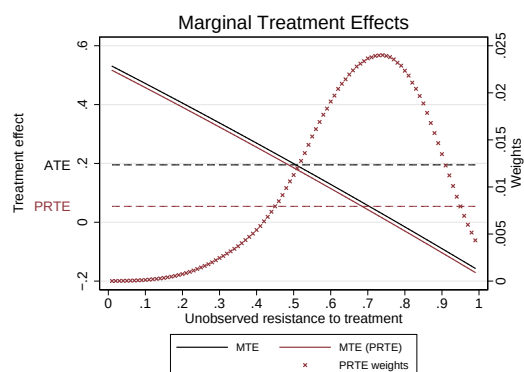
(a) Shifts in propensity scores from alternative policy



(b) Reading



(c) Math



(d) English

Figure 12: Policy relevant treatment effects for a further child care expansion

Note: Figure illustrates the Policy Relevant Treatment Effects of a further expansion of child care. Assigning the coverage rates in their municipality in 2009 to children in the sample generates shifts in the propensity scores as illustrated by the top left figure. These shifts in propensity scores generate the PRTEs illustrated in the three other figures.

Table 5: Treatment effect parameters

	Parameter	Reading	Math	English
ATE	Average treatment effect	0.189 (0.242)	0.355 (0.257)	0.195 (0.307)
ATT	Average treatment effect on the treated	0.199 (0.406)	0.441 (0.402)	0.290 (0.478)
ATUT	Average treatment effect on the untreated	0.166 (0.201)	0.173 (0.212)	-0.00441 (0.188)
LATE	Local average treatment effect	-0.0880 (0.134)	-0.0291 (0.146)	-0.0885 (0.149)
PRTE	Policy relevant treatment effect	0.165*	0.181*	0.0539

Note: Treatment effect parameters calculated as weighted averages over the particular MTE for the population of interest. Standard errors in parentheses, clustered at the municipality level. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

estimation sample ends because the kids exposed to the reform in the later years have not yet taken the tests. In fact, the coverage rates continued to increase, attaining a maximum around 80% coverage for 1- and 2-year olds in 2009-2010. I therefore consider the actual continued rollout of the reform, and predict the effects of the rollout of the reform in 2008-2009 on the test scores of the children induced to enroll due to this further expansion. To this end, I assign the coverage rate in the municipality in 2009 to every child in the sample. Using the first stage, I can predict how the distribution of propensity scores will change relative to the baseline policy that existed in 2004–2007. The shift in propensity scores is depicted in the top left panel of figure 12 - we see that the continued expansion is predicted to shift the propensity scores to the right.

Using these shifts, I calculate the weights that the PRTE put on each individual and each point on the U_D distribution. In a nutshell, the PRTE puts more weight on people who have a large shift in their propensity scores due to the policy shift: These are more likely to be compliers to the policy change.

The compliers to this policy change seems to be relatively similar to the average child in terms of observables: The MTE curves are very similar for math and English, and only slightly lower for reading. The weights, however, oversample children from the upper part of the U_D distribution: People with U_D in the range .5 to .8 are far more likely to be shifted into treatment by this hypothetical policy change. This is why we find a Policy Relevant Treatment Effect of .17, .18 and .06 when calculating the PRTE as a weighted average over the MTE curve: The policy induces children with higher than average resistance to start care, and because of the downward sloping MTE curves, these children have below

average treatment effects. This exercise illustrate how further expansion is likely to yield smaller gains despite that the children already treated had large gains: The reform have already served the children with the highest treatment effects, and so future expansion is expected to have lower impact.

7 Robustness

Having documented these patterns in the marginal treatment effects, I move to consider a range of specification- and robustness checks for alternative explanations for these findings.

First stage model

In the baseline specification, I use a linear probability model for estimating the propensity score. This generates probabilities of treatment outside $(0, 1)$, but has the advantage of not imposing function form and using that to estimate the propensity. Common alternatives are the probit and logit models. To see whether the manual adjustment of the propensity scores above 1 or below 0 could be driving my result, I estimate the first stage using a probit or logit model instead. In the top panel of figure 13, I show the common support plots for the three first stage models. The linear probability model has bunching at 1 and to a lesser extent at 0 while this is not the case for the probit and logit models. They have distribution of propensity scores that are skewed to the left for the treated sample. Otherwise, the logit and probit look very similar.

In the bottom panel of figure 13, I plot the resulting MTEs using the predicted propensity scores from these three models. For math and English, the MTE estimates using the different first stage models is very similar. For reading, the probit and logit estimates show much more of a nonlinearity. I take this as some evidence that the functional form of the first stage is not driving the shape of the MTE curve, particularly not for math and reading.

Parametric and semiparametric estimates

To investigate the particular choice of the parametric polynomial as my baseline result, I want to explore the shape of the estimated MTE curves using other parametric and semiparametric MTE estimators. In figure 14, I therefore show the estimated MTEs using four different methods: The parametric normal, the parametric polynomial, the semiparametric polynomial and the fully semiparametric model. Note that when estimating the parametric polynomial model, it is necessary to use a probit for the first stage, which might affect any differences in addition to the estimation method itself.

When estimating semiparametric models, common support is crucial: We can only identify the MTEs at points of the U_D distribution where we have common support of

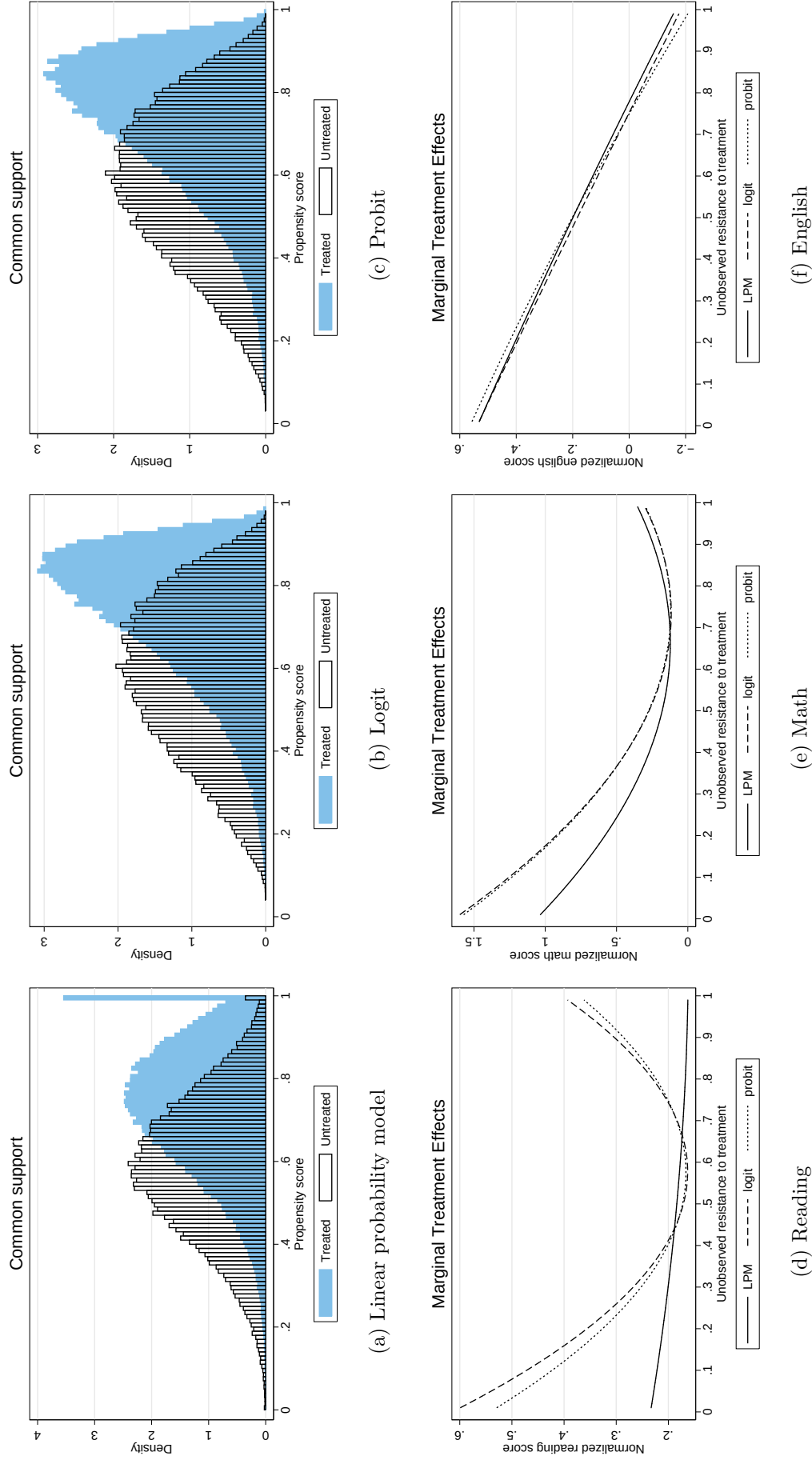


Figure 13: Alternative first stage models

Note: Top panel shows the common support graphs for the three alternative probability models. Bottom panel shows the marginal treatment effect curves using the resulting propensity scores.

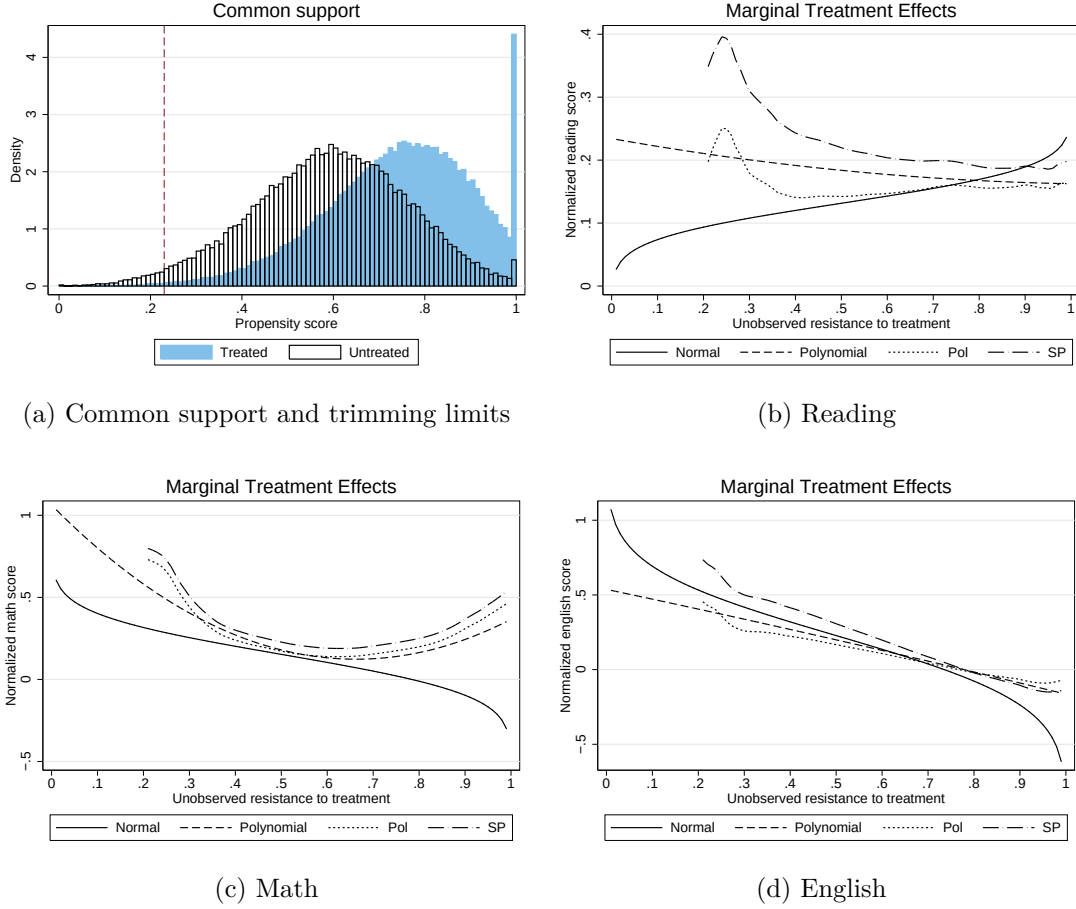


Figure 14: Parametric and semiparametric estimates of the MTE

propensity score in the treated and untreated samples. I therefore trim the sample so that the 0.25% of both the treated and untreated sample that belong to the thinnest tail of the propensity score distribution is removed, and estimate the MTE only at the points where there is overlapping support in both samples after trimming. The trimming limit and the common support plot using this trimming rule is depicted in the top left panel of figure 14.

For all three outcomes, we see that the parametric polynomial model is a relatively good fit to the semiparametric MTE estimate at the points where we have support. The parametric normal, differ considerable, particularly for reading, where it is upward sloping.

Degree of polynomial

In the baseline specification, I use a polynomial of degree 2 to model the $k(u)$ function. This restricts the MTE function to be a quadratic function in the unobserved resistance to treatment, which is clearly arbitrary. I therefore investigate the sensitivity of the MTE to this choice using polynomials of degree 1 to 5, as well as the semiparametric polynomial MTE as a guideline. These results are reported in figure 15.

Starting with reading, we see that most of the estimates line up relatively well. All are relatively flat or downward sloping, except for some strong negative trend at the beginning of the fifth order specification. Remember, however, that common support is very sparse up until about .2, meaning we should be cautious in trusting estimates in the range below this. We find a similar pattern for math and English, with downward sloping MTEs for most specifications, particularly in the bulk of the data. I conclude that a second order polynomial seems to be a fair choice for baseline specification.

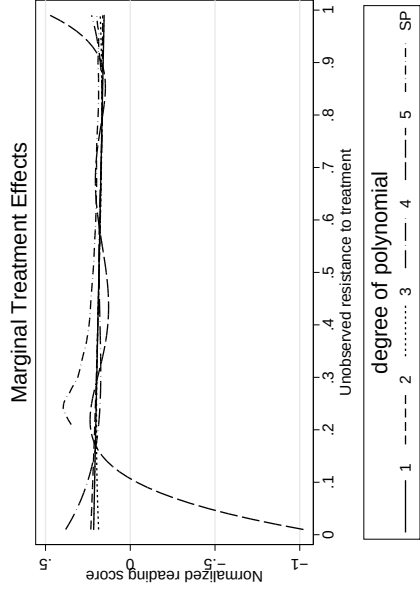
Municipality-specific linear time trends

Despite the relatively unpredictable expansion of care that followed the reform and the control for local aggregate demand, we might worry about omitted variable bias. In particular, if underlying trends in future test scores in each municipality correlate with expansions of care, my estimates of the effect of care might reflect these heterogeneous trends rather than true causal impacts. To control for this, I condition on municipality-specific linear time trends by controlling for the interaction between a continuous time measure and the municipality-fixed effects when estimating the MTE. I allow these to differ by treatment status like all the other covariates. The results of these exercises are plotted in figure 16. Although results are very imprecise, the MTE curves using this specification is relatively similar to the baseline, and if anything, the downward sloping trend is stronger.

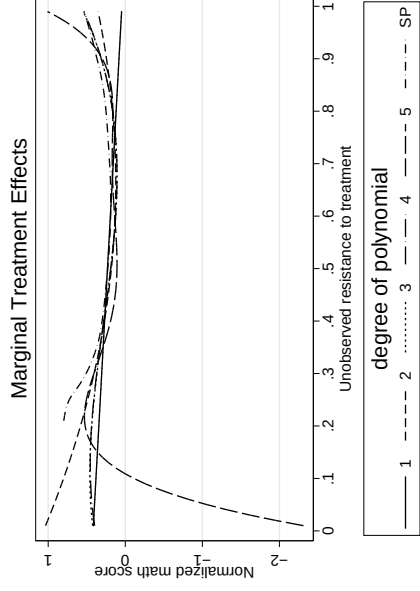
8 Conclusion

Expansion or introduction of universal child care programs, rather than targeted ones, are attracting increasing attention from policymakers worldwide. Often, these are claimed to equalize differences in early skills because children from disadvantaged families benefit more. This paper has investigated the pattern of selection into early child care as well as the heterogeneity in the treatment effect of early child care on school performance at age 10, and in particular this claim. Identification comes from a large-scale reform of child care for toddlers in Norway in 2002 that expanded care unequally across municipalities and over time.

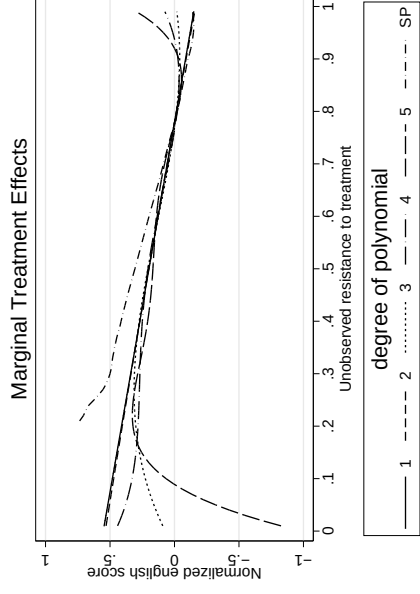
I first document the patterns of selection into early care, and unsurprisingly find higher enrollment among children from advantaged households along most dimensions. Using the Marginal Treatment Effects framework, I then document how the treatment effect varies across a) observable covariates and b) unobserved resistance to treatment, often interpreted as taste or preferences. Despite the relatively long time between treatment at ages 1-2 and measuring of skills at age 10, I find some effects of early child care on school



(b) Reading

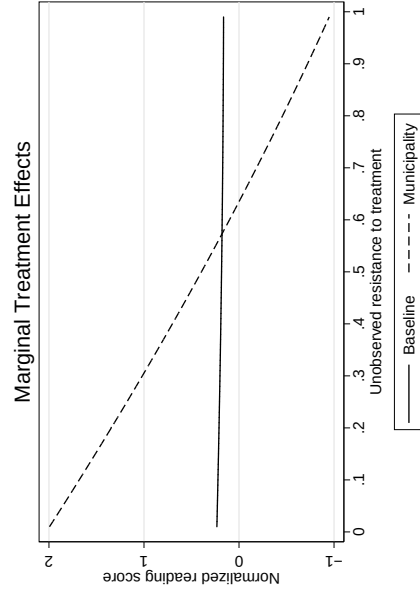


(c) Math

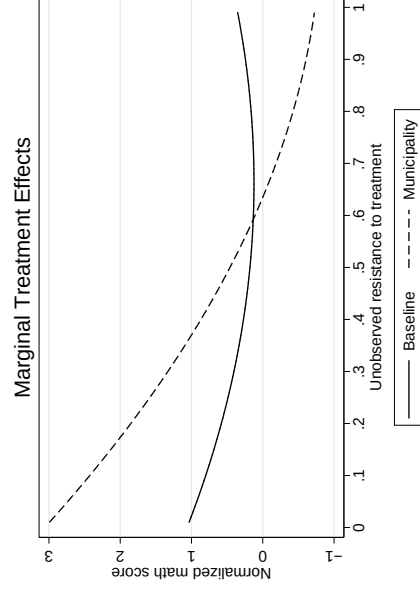


(d) English

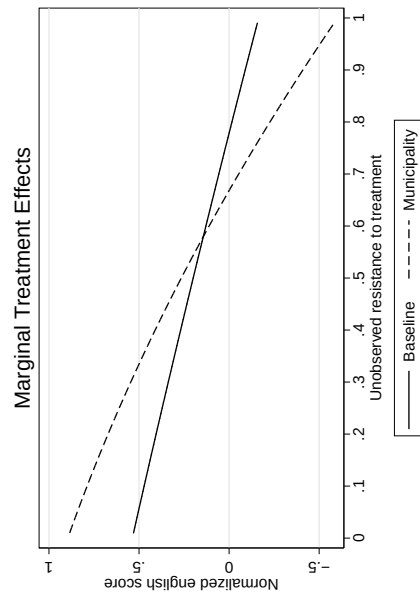
Figure 15: Degree of polynomial



(b) Reading



(c) Math



(d) English

Figure 16: Controlling for municipality-specific linear time trends

performance. Although imprecisely estimated, average treatment effects are relatively large at .2 to .44 standard deviations increase in test scores in reading, math and English. There is some evidence of positive selection on observable gains coming from the fact that children with observables that give them larger treatment effects are more likely to select into treatment.

Furthermore, the MTE curves are downwards sloping, even significantly so for math, indicating that the children who are most likely to choose or be chosen for treatment has a higher treatment effect. This is consistent with a traditional Roy model where we find positive selection on unobserved gains: Children with large unobserved gains are more likely to select care precisely because they benefit more. This is an indication that the decision to enroll could be taken at least with partial knowledge of the unobserved gains.

These results stand in contrast to the traditional instrumental variables results, that indicate no or even negative treatment effects for the complier group. This is an indication of essential heterogeneity in the response to early care, and a sign that we should be careful in interpreting local average treatment effects: The compliers might be very particular populations. By calculating the policy-relevant treatment effect, I use the MTE model to predict the effect of the continued expansion of care that happened after the end of my sample period. Because of the downward-sloping MTE and the positive selection, further expansion is expected to give smaller treatment effect than the initial reform, precisely because the children who have the most to gain are already served.

This analysis has important policy implications for policymakers considering introductions or expansions of universal child care systems. There are some potential for equalizing initial differences in endowments using universal intervention, but because of heterogeneity in the treatment effect, the efficiency of an expansion depends crucially on the selection into treatment. Further research should seek to explore this, particularly the nature of selection under rationing of treatment.

References

- ALMOND, D. AND J. CURRIE (2011): “Human Capital Development before Age Five,” in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 4, Part B of *Handbook of Labor Economics*, 1315 – 1486.
- ANDRESEN, M. E. (2016): “Exploring Marginal Treatment Effects: Flexible Estimation using Stata,” Working paper.
- ASPLAN VIAK (2004-2010): “Analyse av barnehagestatistikk pr. 20. september [Analysis of child care statistics by September 20th],” Tech. rep., Ministry of Education.
- (2007): “Sluttstatus i forhold til målet om full barnehagedekning 2007 [Final status for the goal of full child care coverage, 2007],” Tech. rep., Ministry of Education.
- BJÖRKLUND, A. AND R. MOFFITT (1987): “The Estimation of Wage Gains and Welfare Gains in Self-selection,” *The Review of Economics and Statistics*, 69, 42–49.
- BRINCH, C., M. MOGSTAD, AND M. WISWALL (2015): “Beyond LATE with a discrete instrument,” Working paper.
- CARNEIRO, P., J. J. HECKMAN, AND E. J. VYTLACIL (2011): “Estimating Marginal Returns to Education,” *American Economic Review*, 101, 2754–81.
- CARNEIRO, P. AND S. LEE (2009): “Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality,” *Journal of Econometrics*, 149, 191–208.
- CASCIO, E. U. AND D. W. SCHANZENBACH (2013): “The Impacts of Expanding Access to High-Quality Preschool Education,” Working Paper 19735, National Bureau of Economic Research.
- CORNELISSEN, T., C. DUSTMANN, A. RAUTE, AND U. SCHÖNBERG (2015): “Who benefits from universal child care? Estimating marginal returns to early child care attendance,” Working paper.
- (2016): “Who benefits from universal child care? Estimating marginal returns to early child care attendance,” Forthcoming in *Journal of Political Economy*.
- CUNHA, F. AND J. HECKMAN (2007): “The Technology of Skill Formation,” *American Economic Review*, 97, 31–47.
- CURRIE, J. (2001): “Early Childhood Education Programs,” *Journal of Economic Perspectives*, 15, 213–238.
- CURRIE, J. AND D. THOMAS (1995): “Does Head Start Make a Difference?” *American Economic Review*, 85, 341–64.
- FELFE, C. AND R. LALIVE (2015): “Does Early Child Care Affect Children’s Development?” Working paper.
- FITZPATRICK, M. D. (2008): “Starting School at Four: The Effect of Universal Pre-Kindergarten on Children’s Academic Achievement,” *The B.E. Journal of Economic Analysis & Policy*, 8, 46.

- HAVNES, T. AND M. MOGSTAD (2011): “Money for nothing? Universal child care and maternal employment,” *Journal of Public Economics*, 95, 1455 – 1465, special Issue: International Seminar for Public Economics on Normative Tax Theory.
- HECKMAN, J. AND E. VYTLACIL (2001): “Local instrumental variables,” in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics. Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell, Cambridge Univ. Press, New York, 1–46.
- HECKMAN, J. J., S. URZUA, AND E. VYTLACIL (2006): “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *The Review of Economics and Statistics*, 88, 389–432.
- HECKMAN, J. J. AND E. VYTLACIL (2005): “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669–738.
- HECKMAN, J. J. AND E. J. VYTLACIL (1999): “Local instrumental variables and latent variable models for identifying and bounding treatment effects,” *Proceedings of the National Academy of Sciences of the United States of America*, 96(8), 4730–4734.
- (2007): “Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments,” Elsevier, vol. 6, Part B of *Handbook of Econometrics*, 4875 – 5143.
- IMBENS, G. W. AND J. D. ANGRIST (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, pp. 467–475.
- KLINE, P. AND C. WALTERS (2015): “Evaluating Public Programs with Close Substitutes: The Case of Head Start,” Working Paper 21658, National Bureau of Economic Research.
- MAESTAS, N., K. J. MULLEN, AND A. STRAND (2013): “Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt,” *American Economic Review*, 103, 1797–1829.
- MINISTRY OF EDUCATION (2008-2009): “Kvalitet i barnehagen [White Paper no. 41],” Available at <http://www.regjeringen.no/nb/dep/kd/dok/regpubl/stmeld/2008-2009/stmeld-nr-41-2008-2009-.html>.
- MINISTRY OF EDUCATION AND RESEARCH (1998): “OECD - Thematic Review of Early Childhood Education and Care Policy,” Available at www.oecd.org/dataoecd/48/53/2476185.pdf.
- (2002-2003): “Barnehagetilbud til alle - økonomi, mangfold og valgfrihet White Paper no. 24,” Available at <http://www.regjeringen.no/nb/dep/kd/dok/regpubl/stmeld/20022003/stmeld-nr-24-2002-2003-.html>.
- VYTLACIL, E. (2002): “Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” *Econometrica*, 70, 331–341.