Long-Term Distributional Effects of Early Tracking: Evidence from a Comprehensive School Reform in Finland*

Miikka Rokkanen†

September 23, 2011

Abstract

In the 1970s the tracking of students into academic and non-academic tracks was postponed in Finland from age 11 to age 16 by the introduction of a uniform 9-year comprehensive school. In this paper I utilize the gradual implementation of the comprehensive school reform across the nation to capture the causal effect of early tracking on the distribution of cognitive skills and earnings substantially later in life. I estimate these effects using register data on Finnish men born in 1962-1966 and the changes-in-changes method by Athey and Imbens (2006a). According to my results, early tracking had no effects on the distribution of cognitive skills. However, it increased earnings in the highest quartile of the distribution while the bottom half of the earnings distribution was instead unaffected by early tracking. I also find substantial heterogeneity in the earnings effects along parental background. Most of the effects seem to come from men with educated or high earning parents. For cognitive skills there is instead no evidence of heterogeneity in the effects along these dimensions.

*I am grateful to Roope Uusitalo and Tuomas Pekkarinen for their help with the data used in this paper. I would also like to thank Joshua Angrist, David Autor, Benjamin Feigenberg, Jerry Hausman, Ossi Korkeamäki, Juan Passadore, Parag Pathak, Matti Sarvimäki, Annalisa Scognamiglio and the participants at MIT’s labor and econometrics lunches for their helpful comments and suggestions.

†Department of Economics, Massachusetts Institute of Technology. Email: rokkanen@mit.edu.
1 Introduction

The choice between selective and comprehensive school systems is an important decision in education policy. During the post-war period the general trend among OECD countries has been towards providing more uniform education to individuals at least until the end of compulsory schooling. However, there is still substantial heterogeneity across countries in the age at which students are tracked into academic and non-academic tracks. For instance, in Austria and Germany the tracking of students happens already at age 10 whereas in countries like the UK and the US tracking does not take place before age 16.

On the other hand, even countries with comprehensive school systems are often characterized by more implicit tracking of students through school choice policies and the existence of private schools with high tuition and public schools with selective enrollment. These features of the education system can be seen as mechanisms that potentially increase the segregation of students according to ability and family background. Moreover, it is relatively common that students are tracked within schools based on prior achievement.

The arguments for and against tracking are typically related to a trade-off between efficiency and equity. According to one view, tracking can increase the performance of all students as homogenous groups allow educators to target teaching and resources more effectively (Duflo, Dupas, and Kremer, 2011). On the other hand, tracking might hurt low ability students but nevertheless increase the average performance in case these losses are outweighed by benefits for high ability students. Opponents of tracking are instead worried that it increases inequality in the economy by widening the performance gap between high and low ability students without necessarily having any effects on average performance.\footnote{Theoretical discussion on the consequences of tracking can be found in Benabou (1996), Lazear (2001), Epple, Newlon, and Romano (2002) and Brunello, Giannini, and Ariga (2007). See also Eckstein and Zilcha (1994), De Fraja (2002), Fleurbaey, Gary-Bobo, and Maguain (2002), Hanushek, Leung, and Yilmaz (2003) and Ben Mimoun (2005) for theoretical analysis related to education policy and redistribution.}

In this paper I shed light on the long-term distributional consequences of a school system where students are tracked into academic and non-academic tracks at a relatively early age. For this purpose I utilize a comprehensive school reform that postponed the tracking of students in Finland from age 11 to age 16. The reform was gradually implemented across the nation during the period 1972-1977 which generates a promising quasi-experimental setting for studying the causal effects of early tracking. As a result of the regional implementation plan two school systems coexisted in Finland for approximately a decade, causing variation across cohorts and municipalities in whether a student went through a tracked or a comprehensive school system during primary education.

My identification and estimation strategy is based on the changes-in-changes method by Athey and Imbens (2006a). This is a generalization of the widely-used difference-in-differences method that can be used to nonparametrically identify and estimate the entire counterfactual distribution for the outcome of interest. This allows me to look at the effect effects of early tracking more broadly by comparing the quantiles of the potential outcome distributions under the tracked and
comprehensive school systems.

I use in this paper register data covering almost the entire population of Finnish men born in 1962-1966. To get a long-term perspective on the effects of early tracking I focus on two outcomes observed substantially later in life. The first outcome I look at is the performance in the Finnish Army Basic Skills test which is a test designed to measure cognitive skills and taken by conscripts in the Army typically around age 20. In addition, I look at the annual earnings of the individuals in 1996-2003 when they are, depending on the cohort, around age 30-41.

I find no effects of early tracking on Army test scores, but it seems to instead widen the earnings distribution by increasing earnings especially in the highest quartile of the distribution. Below the median the distribution is instead unaffected by the school system. I also look at the effects of early tracking separately for subsamples based on parental education and earnings. These findings suggest that the effects found for earnings are mainly coming from individuals with educated or high earning parents. For Army test scores there is instead no evidence of heterogeneity in the effects of early tracking along these dimensions. Finally, I conduct a number of specification and robustness checks that provide strong support for the validity of my findings.

Several papers have previously looked at the effects of tracking on student achievement. For instance, the papers by Duflo, Dupas, and Kremer (2011), Zimmer (2003) and Lefgren (2004) report positive effects of tracking within schools in Kenya and the US. The US evidence is mixed, however, as Betts and Shkolnik (2000) and Figlio and Page (2002) find no effects of within-school tracking on test scores whereas the results by Argys, Rees, and Brewer (1996) suggest this to benefit high-achieving students at the expense of low-achieving students.

Galindo-Rueda and Vignoles (2004), Bonhomme and Sauder (2011) and Guyon, Maurin, and McNally (2010) have instead studied the effects of tracking students across schools on student achievement in the UK finding again somewhat mixed results. If anything, these studies suggest selective school system to benefit high-achieving students. Muhlenweg (2007) finds also similar results for Germany where tracking across schools seems to widen the achievement distribution. Moreover, in contrast to my findings, Pekkarinen, Uusitalo, and Kerr (2009a) report that the Finnish comprehensive school reform increased Army tests scores among individuals with uneducated parents.

Some studies have also looked at cross-country evidence on the impacts of tracking on student achievement. Amermuller (2005) and Hanushek and Woessmann (2006) have reported tracking to increase the role of family background and increase inequality in student achievement, but these results seem highly controversial due to the conflicting evidence by Waldinger (2007). On the other hand, the results by Ariga and Brunello (2007) suggest years spent in a tracked school system to increase average test scores.

Potentially due to data reason, the effects of tracking on educational attainment and earnings have been studied less in the literature. In the UK setting Galindo-Rueda and Vignoles (2004) have shown a selective school system to increase the educational attainment of high-achieving students whereas Guyon, Maurin, and McNally (2010) report positive effects of increasing acces to a

---

2See also the discussion in Manning and Pischke (2006).
selective track on educational attainment. Malamud and Pop-Eleches (2011) find instead that the postponement of tracking in Romania did not improve access to higher education for disadvantaged children where as Kang, Park, and Lee (2007) report negative effects of tracking on the bottom of the earnings distribution in South Korea.

This paper is also closely related to the literature studying the role of tracking in intergenerational mobility of education and earnings. For instance, Pekkarinen, Uusitalo, and Kerr (2009b) have used the Finnish comprehensive school reform to show evidence on the detrimental effect of tracking on intergenerational mobility, and similar findings have also been reported by Bauer and Riphahn (2006) for Switzerland and by Brunello and Checchi (2007) using a cross-country setting. Pekkarinen (2008) shows instead that the Finnish comprehensive reform increased the gender difference in educational attainment and reduced the gender wage gap. Moreover, Meghir and Palme (2005), Holmlund (2008), Black, Devereux, and Salvanes (2005) and Aakvik, Salvanes, and Vaage (2010) report increases in educational attainment and earnings as well as evidence of increasing intergenerational mobility of education and earnings as a result of the comprehensive school reforms in Sweden and Norway.

The rest of this paper is organized as follows. In section 2 I go through the main features of the pre- and post-reform school systems in Finland as well the implementation of the reform. This is followed by a description of the data and the CIC method in section 3. In section 4 I present my main results as well a number of specification and robustness checks to evaluate the validity of my findings. Finally, section 5 concludes the paper.

2 Finnish Comprehensive School Reform

2.1 Pre- and Post-Reform School Systems in Finland

During the 1970s the Finnish education system went through a dramatic change due to a comprehensive school reform that was gradually implemented across the nation. The reform was mainly motivated by an attempt to enhance social, economic and regional equality by providing equal educational opportunities to all students irrespective of their place of residence and parental background. It followed the footsteps of similar educational reforms that had already taken place in Sweden in the 1950s and Norway in the 1960s.3

Until the 1970s Finnish primary education was characterized by a two-track school system where students were divided into an academic and non-academic track already at a relatively early age. This school system is illustrated in the left column of figure 1. In the two-track school system all students entered primary school at age 7. At age 11, after only four years of uniform education in primary school, students could either apply to a general secondary school or continue in the primary school.

Students were selected into the general secondary school based on primary school grades, teacher

---

3For more details and background on the comprehensive school reform in Finland, see Aho, Pitkänen, and Sahlberg (2006).
assessments and an entrance exam. It was an academic track that lasted for five years and opened up the possibility to continue into upper secondary school for three additional years. At the end of upper secondary school students could take a matriculation exam that made them eligible for applying to university. In the non-academic track, on the hand, students continued in primary school for two years after which most of the students continued to civic school for two or three additional years. After civic school students could continue their education in vocational school, whereas they were not eligible to continue to a upper secondary school. While in the tracked school system compulsory schooling consisted of only the 6-year primary school, in the 1970s also students that were following the non-academic track typically stayed in school for at least 8 or 9 years.

There were considerable differences in the curriculums followed in the academic and non-academic track. For instance, in the academic track it was compulsory to study foreign languages, and students learned more advanced mathematics and science. The purpose of the non-academic track was instead to prepare students for low-skilled occupations and thus the focus in education was more on practical skills relevant for these jobs.

![School system in Finland before and after the comprehensive school reform](image)

Figure 1: School system in Finland before and after the comprehensive school reform

During the 1970s the two-track school system was abolished and replaced with the comprehensive school system illustrated in the right column of figure 1. In this school system all student entered a uniform 9-year comprehensive school at age 7. In comprehensive school the students followed
the same curriculum at the same establishments. This postponed the tracking of students into an academic and non-academic track in Finland to age 16 when students could apply to either upper secondary school or vocational school based on their comprehensive school grades.

The comprehensive school curriculum had a more academic content when compared to primary school and civic school. It resembled to a large extent the general secondary school curriculum due its greater focus on mathematics, science and foreign languages. However, the level of education had to be adjusted to the more heterogeneous student body, and thus the curriculum in the comprehensive school became less demanding than in the general secondary school.

To summarize, the main difference between the pre- and post-reform school systems in Finland was the age at which students were divided into an academic and non-academic track. While in the pre-reform school system students were tracked already at age 11, tracking did not happen until age 16 in the post-reform school system. This exposed students who would have otherwise gone to the non-academic track to a curriculum with a higher academic content whereas it reduced the level of education for student who would have otherwise gone to the academic track.

2.2 Implementation of the Reform

The comprehensive school reform was implemented in Finland gradually across the nation during the period 1972-1977. The adoption of the reform in each municipality followed a regional implementation plan illustrated in figure 2. The first municipalities that adopt the comprehensive school system in 1972 were predominantly located in northern Finland. In 1973 and 1974 the comprehensive school system was then adopted in northeastern and northwestern parts of Finland. In 1975 and 1976 the reform spread to the southeast and the southwest. Finally, Helsinki and its surrounding municipalities were the last ones to adopt the comprehensive school system in 1977.

Within each municipality the adoption of the comprehensive school system was done as follows. In the fall of the year a municipality implemented the reform grades 1-5 immediately started in the comprehensive school system. Students who were already starting 6th grade in the fall of that year completed their primary education instead in the tracked school system. Thus, it took approximately four years for a given municipality to fully adopt the comprehensive school system in the sense that all of the students attending grades 1-9 in the municipality were covered by the new system.

The gradual implementation of the comprehensive school reform in Finland offers a promising quasi-experimental setting for studying the causal effects of early tracking. There is variation across both cohorts and municipalities in school system attended by individuals around the time of the reform, providing the basis for the identification and estimation strategy used in this paper. However, there are some caveats that should be noted here. First, some municipalities already had experimental comprehensive schools before the actual reform took place. Second, some municipalities had founded a few years before the comprehensive school reform general secondary schools that enrolled almost all of the relevant cohorts in the municipality. Third, ability grouping was retained in mathematics and foreign languages within the comprehensive school system until 1985. While it
is difficult to evaluate the magnitude of the effect of these factors on my estimates, one would expect all of them to go against finding any effect of early tracking and thus to attenuate my findings.

3 Empirical Strategy

3.1 Data

The data for this paper comes from two sources: the Finnish Army database and the Statistics Finland longitudinal census file. The sample I use is based on the Army database and covers the full population of men who were born between 1962-1966 and started their military service after January 1982. For these individuals I observe the date and result of their Finnish Army Basic Skills test. Statistics Finland has then linked various other information to this data from the census file based on personal identification numbers.
The Finnish Army Basic Skills test is a test designed to measure general abilities and used by the Finnish Army to select conscripts to officer training. The test was originally created in 1955 and re-designed in 1981, but during the years observed in the data it remained unchanged. The test consists of three sections (verbal, arithmetic and logical reasoning) that each consist of 40 multiple choice questions. The verbal and arithmetic reasoning sections measure skills primarily taught in school whereas the logical reasoning section should be less affected by schooling. The first two sections ask the subject, for instance, to choose synonyms or antonyms for given words, identify group memberships of words, complete number series, solve mathematical problems and to identify similar relationships between pairs of words and numbers. The logical reasoning section is instead a standard intelligence test based on the Raven’s progressive matrices.

The Statistics Finland longitudinal census file contains data on the entire population living in Finland and is mainly based on administrative registers. The data has only a few people with missing information, and the main reasons for not being included in the data are residence abroad and death. From the census file I have information on, for instance, year of birth, place of residence in 1970, 1975 and 1980 as well as on education, wage income and income from self-employment in 1980, 1985, 1990 and 1995-2003. In addition, I observe the completed education of both parents in 1980 as well as their total taxable income in 1970, 1975, 1980, 1985 and 1990.\footnote{In addition to wage income and income from self-employment, total taxable income includes also capital gains and various social benefits.} All of the income data are based on individual tax records and top-coded to ensure that individuals cannot be identified from the data. However, this affects only less than 1% of the observations for any given income variable.

Statistics Finland does not release the data to researchers with municipality-level information on place of residence although this information is included in the original census file. The place of residence in 1970, 1975 and 1980 is coded in the data instead according to the year in which a given municipality implemented the comprehensive school reform as illustrated in figure 2. Together with year of birth this information can be used to determine whether an individual attended the tracked or comprehensive school system. Unfortunately, this approach only works for individuals who did not move during the relevant years between municipalities with different school systems in place for their cohort. I exclude the movers from the data used in the analysis below.

The main outcomes I focus on in this paper are the average score in the three sections of the Finnish Army Basic Skills test and the average annual earnings in 1996-2003. I standardize the average test score to have mean 0 and variance 1 within a test year and define annual earnings as the sum of wage income and income from self-employment. I also look at subsamples of individuals based on parental education and income. For parental education I use the highest degree among mother and father, and for parental income I use the average total taxable income of mother and father in 1970, 1975, 1980, 1985 and 1990. All income data is inflated/deflated to the prices in year 2000 using the cost of living index.

Since the military service is mandatory for men in Finland, the data covers considerably well the
population of Finnish men born in 1962-1966. It is, however, subject to some sources of selectivity. For instance, it is possible to enter the military service as a volunteer already at age 17 which means that some of the men in the oldest cohorts started their military service before 1982 and are thus not included in the data. In addition, it is possible to be exempted from the military service due to religious or ethical conviction or severe health conditions. Fortunately, in the 1980s the use of these exemptions was less common than what it is today. Based on a comparison to the 1984 population census the data covers around 85% Finnish men born in 1962-1966.

Table 1 illustrates how the exposure to comprehensive school reform is defined according to year of birth and place of residence as well as how the data is distributed across the cohort-region cells. We can see that, for instance, in regions that implemented the reform in 1972 or 1973 all cohorts born in 1962-1966 attended the comprehensive school system while in the region that implement the reform in 1977 this was true only for the cohort born in 1966. Altogether 64% of men born in 1962-1966 were exposed to the comprehensive school reform.

Table 1: Exposure to comprehensive school reform by cohort and region

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1962</td>
<td></td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[2,667]</td>
<td>[3,944]</td>
<td>[5,754]</td>
<td>[5,523]</td>
<td>[5,805]</td>
<td>[3,093]</td>
</tr>
<tr>
<td>1963</td>
<td></td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[2,987]</td>
<td>[4,418]</td>
<td>[6,444]</td>
<td>[6,583]</td>
<td>[6,619]</td>
<td>[3,769]</td>
</tr>
<tr>
<td>1964</td>
<td></td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[2,910]</td>
<td>[4,357]</td>
<td>[6,299]</td>
<td>[6,550]</td>
<td>[6,809]</td>
<td>[4,055]</td>
</tr>
<tr>
<td>1965</td>
<td></td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[2,750]</td>
<td>[4,097]</td>
<td>[6,083]</td>
<td>[6,400]</td>
<td>[6,372]</td>
<td>[4,000]</td>
</tr>
<tr>
<td>1966</td>
<td></td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[2,219]</td>
<td>[3,353]</td>
<td>[5,197]</td>
<td>[5,673]</td>
<td>[5,632]</td>
<td>[3,681]</td>
</tr>
</tbody>
</table>

Notes: The bordered area indicates cells that were exposed to the comprehensive school reform. Years spent in comprehensive school reported in each cell. Number of observations with non-missing outcome information in squared brackets.

Table 2 reports descriptive statistics of the data by the school system attended. According to these numbers, both test scores and earnings are, on average, higher for the population who attended the tracked school system. In addition, the dispersion of earnings is higher for this group whereas there is not much difference in the dispersion of test scores. It also seems that the population who attended the tracked school system have, on average, more educated and higher earning parents. However, these simple comparisons are likely to be confounded by the age and regional compositions of these two groups. For instance, while not shown here, the mean differences observed for parental earnings and education become negligible and statistically insignificant once one controls for cohort and region fixed effects.
Table 2: Descriptive statistics by the school system attended

<table>
<thead>
<tr>
<th>Outcome information:</th>
<th>Tracked</th>
<th>Comprehensive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
</tr>
<tr>
<td>Test score</td>
<td>51,359</td>
<td>0.063</td>
</tr>
<tr>
<td>Earnings</td>
<td>50,912</td>
<td>25,157</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parental background:</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Educated parents</td>
<td>45,472</td>
<td>0.550</td>
</tr>
<tr>
<td>Parental earnings</td>
<td>45,472</td>
<td>31,896</td>
</tr>
</tbody>
</table>


3.2 Estimation Method

I utilize in this paper the gradual implementation of the comprehensive school reform in Finland to capture the causal effect of early tracking on the distribution of test scores and earnings. For the identification and estimation of these effects I use the changes-in-changes method (CIC) by Athey and Imbens (2006a). CIC is a generalization of the difference-in-differences method (DID) which has become extremely popular in the literature on evaluating the effects of policy changes.\(^5\) CIC has several nice features compared to DID. It, for instance, accommodates the possibility of systematic heterogeneity in the treatment effect and is invariant to arbitrary monotonic transformations of the outcome. Most importantly for this paper, under the assumptions of CIC, one can nonparametrically identify and estimate the entire counterfactual distribution for the outcome of interest.

In this section I will briefly cover the identification and estimation of the average treatment effect and quantile treatment effects using CIC in the context of the Finnish comprehensive school reform.\(^6\) I will estimate these parameters for the full population and thus consider the identification of counterfactual distributions for both the treatment and control groups. It should be noted that the average treatment effect and quantile treatment effects are point-identified only for a continuous outcome. I focus here solely on this case as the outcomes I will be looking at in this paper, test scores and earnings, can be thought of as roughly continuous.

Suppose we have a random sample \(i = 1, \ldots, N\) of individuals from a target population, and that for each individual \(i\) we observe data on the triple \((Y_i, G_i, T_i)\). Let \(G_i\) denote the region in which individual \(i\) was living during the comprehensive school reform, \(T_i\) the year of birth of individual \(i\) and \(Y_i\) the observed outcome for individual \(i\). Furthermore, let \(Y_i^0\) and \(Y_i^1\) be the potential outcomes for individual \(i\) had he attended the comprehensive school system and the tracked school

---

\(^5\)See, for instance, Heckman, Lalonde, and Smith (1999) and Imbens and Wooldridge (2009) and the references therein.

\(^6\)See Athey and Imbens (2006a) for more details on the CIC framework.
system. Letting $D_i$ be an indicator for attending the tracked school system the observed outcome for individual $i$ can be expressed as

$$ Y_i = (1 - D_i) Y_i^0 + D_i Y_i^1. \quad (1) $$

To identify the counterfactual distributions of the outcome for the individuals I make the following assumptions. First, I assume that we can write the outcome for individual $i$ had he attended the comprehensive school system as $Y_i^0 = h^0(U_i^0, T_i)$ and similarly the outcome for individual $i$ had he attended the tracked school system as $Y_i^1 = h^1(U_i^1, T_i)$ where $U_i^0$ and $U_i^1$ represent the unobserved characteristics of individual $i$ in the two counterfactual states. Furthermore, I assume that $h^0$ and $h^1$ are both strictly increasing functions in $U_i^0$ and $U_i^1$ for all $T_i$. Finally, I assume that $U_i^0$ and $U_i^1$ are identically distributed and independent of $T_i$ conditional on $G_i$ as well as that the support of $U_i^0$ and $U_i^1$ conditional on $G_i$ is equal across $G_i$. Otherwise the distributions of $U_i^0$ and $U_i^1$ are allowed to depend arbitrarily on $G_i$.

Under these assumptions, the effect of attending the tracked school system for individual $i$ can be written as

$$ Y_i^1 - Y_i^0 = h^1(U_i^1, T_i) - h^0(U_i^0, T_i). \quad (2) $$

This illustrates the multiple sources of treatment effect heterogeneity embedded in the model. First, differences in $U_i^0$ and $U_i^1$ as well as $T_i$ are allowed to have a direct influence on the response to the treatment. Furthermore, the distribution of treatment effect can be different across $G_i$ due to the potential dependence of the distributions of $U_i^0$ and $U_i^1$ on $G_i$. Note, however, that as the values of $U_i^0$ and $U_i^1$ are not restricted to be equal, I am only able to identify the marginal distributions of the potential outcomes instead of the distribution of the individual-specific treatment effects.

To ease the notational burden, I will introduce some additional notation. Let $Y_{i,g,t}^0 \overset{d}{\sim} Y_i^0 \mid G_i = g, T_i = t, Y_{i,g,t}^1 \overset{d}{\sim} Y_i^1 \mid G_i = g, T_i = t$ and $Y_{i,g,t} \overset{d}{\sim} Y_i \mid G_i = g, T_i = t$ where $\overset{d}{\sim}$ stands for “is distributed as”. Furthermore, let $F_{Y_{i,g,t}^0}, F_{Y_{i,g,t}^1}$ and $F_{Y_{i,g,t}}$ denote the corresponding conditional cumulative distribution functions and $F_{Y_{i,g,t}^{-1}}, F_{Y_{i,g,t}^{-1}}$ and $F_{Y_{i,g,t}^{-1}}$ the corresponding conditional quantile functions.

Now let $(g_1, t_1)$ be such that $D_i = 1$ for $(G_i, T_i) = (g_1, t_1)$ and $(g_0, t_0)$ be such that $D_i = 0$ for $(G_i, T_i) = (g_0, t_0)$, $(G_i, T_i) = (g_0, t_1)$ and $(G_i, T_i) = (g_1, t_0)$. That is, I consider a region $g_1$ where cohort $t_1$ attended the tracked school system but cohort $t_0$ attended the comprehensive school system as well as a region $g_0$ where both cohorts $t_0$ and $t_1$ attended the comprehensive school system.

Under the CIC assumptions the distribution of $Y_i^0$ conditional on $(G_i, T_i) = (g_1, t_1)$ is identified and

$$ F_{Y_{i,g_1,t_1}^0}(y) = F_{Y_{i,g_1,t_0}} \left( F_{Y_{i,g_0,t_0}^{-1}} \left( F_{Y_{i,g_0,t_1}}(y) \right) \right). \quad (3) $$
Using this result we can write the average effect of the tracked school system on the outcome conditional on \((G_i, T_i) = (g_1, t_1)\) as
\[
\tau^\mu_{g_1, t_1} = E \left[ Y_{i,g_1,t_1}^1 - Y_{i,g_1,t_1}^0 \right] \\
= E \left[ Y_{i,g_1,t_1} \right] - E \left[ F_{Y_{i,g_1,t_1}^{-1}}^{-1} \left( F_{Y_{i,g_0,t_0}^{-1}} \left( Y_{i,g_1,t_0} \right) \right) \right] 
\]
and the effect of the tracked school system on quantile \(q\) of the outcome distribution conditional on \((G_i, T_i) = (g_1, t_1)\) as
\[
\tau^q_{g_1, t_1} = F_{Y_{i,g_1,t_1}^{-1}}^{-1} (q) - F_{Y_{i,g_1,t_1}^{-1}}^{-1} (q) \\
= F_{Y_{i,g_1,t_1}^{-1}}^{-1} (q) - F_{Y_{i,g_1,t_1}^{-1}}^{-1} \left( F_{Y_{i,g_1,t_1}^{-1}}^{-1} (q) \right). 
\]

Similarly, we can identify the distribution of \(Y_{i}^1\) conditional on \((G_i, T_i) = (g_0, t_1)\) and we have that
\[
F_{Y_{i}^1,g_0,t_1} (y) = F_{Y_{i,g_0,t_0}^{-1}} \left( F_{Y_{i,g_1,t_1}^{-1}} \left( y \right) \right). 
\]

The estimation of these treatment effect parameters can be done in a relatively straightforward fashion by plugging in estimators for the cumulative distribution functions and quantile functions. As estimators for these functions I use
\[
\hat{F}_{Y_{i,g,t}} (y) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} 1 \{ Y_{i,g,t} \leq y \} \\
\hat{F}_{Y_{i,g,t}}^{-1} (y) = \inf \left\{ y : \hat{F}_{Y_{i,g,t}} (y) \geq q \right\} 
\]
where \(N_{g,t}\) denotes the number of observations with \((G_i, T_i) = (g,t)\). As a result, for a given
quadruple \((g_1, t_1, g_0, t_0)\), the estimators for \(\tau_{g_1, t_1}^{\mu}\) and \(\tau_{g_1, t_1}^{q}\) can be written as

\[
\hat{\tau}_{g_1, t_1, g_0, t_0}^{\mu} = \frac{1}{N_{g_1, t_1}} \sum_{i=1}^{N_{g_1, t_1}} Y_{i, g_1, t_1} - \frac{1}{N_{g_1, t_0}} \sum_{i=1}^{N_{g_1, t_0}} \hat{F}_{Y_i, g_0, t_1}^{-1} \left( \hat{F}_{Y_i, g_0, t_0} (Y_{i, g_1, t_0}) \right)
\]

\[
\hat{\tau}_{g_1, t_1, g_0, t_0}^{q} = \hat{F}_{Y_i, g_1, t_1}^{-1} (q) - \hat{F}_{Y_i, g_0, t_1}^{-1} \left( \hat{F}_{Y_i, g_0, t_0} (q) \right)
\]

and similarly the estimators for \(\tau_{g_0, t_1}^{\mu}\) and \(\tau_{g_0, t_1}^{q}\) as

\[
\hat{\tau}_{g_0, t_1, g_1, t_0}^{\mu} = \frac{1}{N_{g_0, t_0}} \sum_{i=1}^{N_{g_0, t_0}} \hat{F}_{Y_i, g_1, t_1}^{-1} \left( \hat{F}_{Y_i, g_0, t_0} (Y_{i, g_1, t_0}) \right) - \frac{1}{N_{g_0, t_1}} \sum_{i=1}^{N_{g_0, t_1}} Y_{i, g_0, t_1}
\]

\[
\hat{\tau}_{g_0, t_1, g_1, t_0}^{q} = \hat{F}_{Y_i, g_1, t_1}^{-1} \left( \hat{F}_{Y_i, g_0, t_0} (q) \right) - \hat{F}_{Y_i, g_0, t_1}^{-1} (q)
\]

As was illustrated in figure 2 and table 1, the implementation of the comprehensive school reform in Finland introduced considerable variation across cohorts and regions in the school system attended by individuals. Thus, I am able to compute \(\hat{\tau}_{g_1, t_1, g_0, t_0}, \hat{\tau}_{g_1, t_1, g_0, t_0}, \hat{\tau}_{g_0, t_1, g_1, t_0}, \) and \(\hat{\tau}_{g_0, t_1, g_1, t_0}\) for multiple quadruples \((g_1, t_1, g_0, t_0)\). To be more exact, for each of the 6 regions I can separately estimate the effect of early tracking for cohorts 1962-1965, and for almost all of these effects I will also have multiple estimates due to different choices of the comparison region and cohort.

Instead of reporting all of these separate estimates I combine them and report an average effect. As the following discussion is not specific to any of the treatment effect parameters considered above, I will use in the rest of this section a generic notation for a treatment effect parameter \(\tau_{g,t}\) and its estimator \(\hat{\tau}_{g,t,g',t'}\). With \((g, t, g', t') = (g_1, t_1, g_0, t_1)\) this will correspond to the effect for the treatment group and with \((g, t, g', t') = (g_0, t_1, g_1, t_1)\) to the effect for the control group. Moreover, I let \(\tau_{g,t}\) to stand for either \(\tau_{g,t}^{\mu}\) or \(\tau_{g,t}^{q}\).

After computing the estimates \(\hat{\tau}_{g,t,g',t'}\) for all of of the possible quadruples \((g, t, g', t')\) I stack them in a \(I \times 1\) vector \(\hat{\tau}_{(g,t,g',t')}\). It can be shown that

\[
\sqrt{N} \left( \hat{\tau}_{(g,t,g',t')} - \tau_{(g,t,g',t')} \right) \xrightarrow{d} N(0, V).
\]

where \(V\) is the asymptotic covariance matrix of \(\sqrt{N} \hat{\tau}_{(g,t,g',t')}\) and \(\tau_{(g,t,g',t')}\) the probability limit of \(\hat{\tau}_{(g,t,g',t')}\). Under the CIC assumptions, \(\tau_{(g,t,g',t')} = A \tau_{(g,t)}\) where \(\tau_{(g,t)}\) is a \(J \times 1\) vector of \(\tau_{g,t}\) for all possible values of \((g_1, t_1)\) and \(A\) is a \(I \times J\) matrix of 0-1 indicators such that \(\tau_{g,t,g',t'} = \tau_{g,t}\) for all quadruples \((g, t, g', t')\).

I then combine the different estimates for each \(\tau_{g,t}\) in a \(J \times 1\) vector \(\hat{\tau}_{(g,t)}\). This can be done efficiently by using the estimator

\[
\hat{\tau}_{(g,t)} = \left( A\hat{V}^+ A \right)^{-1} A\hat{V}^+ \hat{\tau}_{(g,t,g',t')},
\]

where \(\hat{V}\) is a consistent estimator of \(V\) and \(V^+\) denotes the Moore-Penrose generalized inverse of \(V\).
It follows then that
\[
\sqrt{N} \left( \hat{\tau}_{\{g,t\}} - \tau_{\{g,t\}} \right) \overset{d}{\to} N \left( 0, (A'V^+A)^{-1} \right).
\] (17)

Finally, I further combine the estimates \( \hat{\tau}_{\{g,t\}} \) by reporting a single average effect across \((g,t)\). Letting \( \Lambda \) denote a fixed \( J \times 1 \) weighting vector the parameter of interest becomes \( \tau_\Lambda = \Lambda' \tau_{\{g,t\}} \) with a natural estimator
\[
\hat{\tau}_\Lambda = \Lambda' \hat{\tau}_{\{g,t\}}.
\] (18)

satisfying
\[
\sqrt{N} \left( \hat{\tau}_\Lambda - \tau_\Lambda \right) \overset{d}{\to} N \left( 0, \Lambda' (A'V^+A)^{-1} \Lambda \right).
\] (19)

I weight each \( \tau_{g,t} \) according to the sample size \( N_{g,t} \). The final parameter estimates \( \hat{\tau}_\Lambda^\mu \) and \( \hat{\tau}_\Lambda^q \) reported in this paper correspond to the average treatment effect (ATE) and quantile treatment effect (QTE) for cohorts 1962-1965.\(^7\) As there is no variation across regions in the school system attended by the individuals born in 1966, I am not able to identify any treatment effects for this population without making further assumptions.

4 Results

4.1 Main Results

Table 3 reports the ATE estimates for the effects of early tracking on test scores and earnings using DID and CIC. The estimates obtained using these alternative identifying assumptions and estimation procedures are nevertheless remarkably similar. According to these results, early tracking does not seem to have had, on average, any effect on the Army test scores. For earnings the estimated ATE is instead around 350 euros and statistically significant.

Figure 3 shows the QTE estimates for the effect of early tracking on the quantiles of the test score distribution. The estimates fluctuate around 0 and show no indications of any systematic patterns across the quantiles. Only one of the estimates is marginally statistically significant which is, given the number of estimates reported, likely to be an arbitrary finding due to sampling variation. Thus, there seems to be no evidence of early tracking affecting the distribution of Army test scores based on these results.

\(^7\)I estimate \( V \) using bootstrap with 1,000 replications. The simulation study presented in Athey and Imbens (2006b) seems to suggest bootstrapping to outperform the analytical formula provided in Athey and Imbens (2006a).

\(^8\)Note that while \( \hat{\tau}_\Lambda^\mu \) estimates the ATE for cohorts 1962-1965 defined as \( E \left[ Y_{1i} - Y_{0i} \mid 1962 \leq T_i \leq 1965 \right] \), \( \hat{\tau}_\Lambda^q \) does not estimate the QTE for cohorts 1962-1965 defined as \( P_{Y_{1i} \leq T_i \leq 1965 \leq 1965} (q) - P_{Y_{0i} \leq T_i \leq 1965 \leq 1965} (q) \) due to the nonlinearity of the quantiles. I will nevertheless use here the term QTE for simplicity. As the whole counterfactual distribution for each \((g_1, t_1)\) and \((g_0, t_1)\) is identified under the CIC assumptions, it seems that one could identify and estimate also the QTE as defined in this footnote. However, I will leave this potential extension for future work.
Table 3: ATE of early tracking on test scores and earnings

<table>
<thead>
<tr>
<th></th>
<th>DID</th>
<th>CIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test score</td>
<td>-0.003</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Earnings</td>
<td>346**</td>
<td>358***</td>
</tr>
<tr>
<td></td>
<td>(143)</td>
<td>(136)</td>
</tr>
</tbody>
</table>

Notes: This table reports DID and CIC estimates of ATE. Standard errors in parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.

In figure 4 I show instead the QTE estimates for the effect of early tracking on the quantiles of the earnings distribution. The picture arising from these results is considerably different from what was seen above for the test score distribution. The estimates fluctuate mainly around 0 and are statistically insignificant for the bottom half of the distribution. Around the middle of the distribution the estimates are instead at above 200 euros and statistically significant. However, the most substantial effects of early tracking can be seen for the top quartile of the distribution for which the estimates are increasing across the quantiles and highly statistically significant. For instance, the effects found for the 90th and 95th quantiles are estimated to be around 1,000 and 2,400 euros. These results thus suggest that early tracking widens the earnings distribution by increasing earnings especially in the top quartile of the distribution, but it does not seem to have consequences for the bottom half of the distribution.

It is also interesting to see whether there is evidence of differential effects of early tracking on the test score and earnings distributions according to parental background. For this purpose I divide the data into two separate groups based on both parental education and earnings. When dividing the data according to parental education I define an individual to have educated parents if at least one of his parents has completed secondary level education. Uneducated parents are instead defined as parents who have both completed at most primary level education. On the other hand, when dividing the data according to parental earnings I use the median parental earnings in the full sample. I define an individual to have high earning parents if the earnings of his parents are above the median. Low earning parents are instead defined as parents earning less than the median.

Table 4 reports the ATE estimates for the effects of early tracking on test scores and earnings for the subsamples using DID and CIC. Similarly to the full sample results shown in table 3, the estimates obtained using these alternative approaches are extremely close to each other. The ATE estimates are small and statistically insignificant for individuals with low educated and high earning parents. For individuals with educated and high earning parents the estimated effects are instead around 500 euros and statistically significant.

Figure 5 shows the QTE estimates for the effect of early tracking on the quantiles of the test score distribution for these subsamples. While there is some slight variation in the estimated effects across subsamples, the results seem broadly consistent with what was seen above for the full sample.
The estimates fluctuate again around 0 and only a few of them are statistically significant. Thus, the finding of no effects of early tracking on the distribution of test scores seems to hold also for the subsamples based on parental education and earnings.

Figure 6 repeats the subsample analysis for the effects of early tracking on the quantiles of the earnings distribution. These results uncover substantial differences in the effects according to parental background. For individuals with uneducated and low earning parents the estimates remain small and mostly insignificant until the very top of the distribution where the estimates are large but also extremely imprecise. For individuals with educated or high earning parents the estimates are instead positive and statistically significant from around the 60th quantile upwards. While these estimates also become more and more imprecise when reaching the highest quantiles, the overall picture arising from figure 6 is that the effects of early tracking on the earnings distribution are mainly coming from individuals with educated and high earning parents.

4.2 Specification and Robustness Checks

I have conducted a number of specification and robustness checks to verify the findings reported above. I will briefly discuss the results from these checks in this section. The tables and figures referred to in this section can be found in the Appendix. Taken together, the specification and robustness checks seem to provide strong support for the validity of my findings.

As the outcomes I look at in this paper are averages over test sections and earnings in different years, one might wonder whether some interesting effect of early tracking hide behind this averaging.
Figure 4: QTE estimates for the effect of early tracking on the quantiles of the earnings distribution

For instance, it could be the case that early tracking affects the performance in the three sections of the Finnish Army Basic Skills tests but that these effects are not captured by the average score. In figures 7, 8 and 9 I address this issue by showing the QTE estimates separately for the standardized score in each test section as well as for earnings in years 1996-2003.

The separate results for the math, verbal and logical reasoning sections of the Finnish Army Basic Skills test shown in figure 7 seem broadly consistent with what I found for the average test score. There is some systematic variation across different test sections as the the QTE estimates are mainly positive for math score and negative for verbal score. However, for almost all of the quantiles the estimates are small in magnitude and statistically insignificant. Similarly, the separate estimations for annual earnings in 1996-2003 shown in figures 8 and 9 reveal no substantial differences relative to the main results shown above. There is of course some year-to-year variation in the QTE estimates, but the findings are qualitatively similar to what was seen for the average annual earnings in 1996-2003.

Given the considerable variation across cohorts and regions in the school system attended, I am able to test the validity of the underlying assumptions of the CIC model. As was mentioned in section 3.2, all the estimators $\hat{\tau}_{g,t,g',t'}$ are consistent for $\tau_{g,t}$ under the null hypothesis ($H_0$) that the CIC assumptions are satisfied. If the model is instead misspecified, they may converge to different probability limits. Thus, a misspecification test can be based on the fact that under $H_0$

\[ N \left( \tilde{\tau}_{g,t,g',t'} - A\tilde{\tau}_{(g,t)} \right) \tilde{V}^+ \left( \tilde{\tau}_{g,t,g',t'} - A\tilde{\tau}_{(g,t)} \right) \xrightarrow{d} \chi^2 (R - K) \]  

(20)
where \( R \) denotes the rank of \( M \). This test has power against a number of violations of the CIC assumptions. For instance, one such violation would be the presence of additive random effects on the cohort-region level. Unlike in the DID model where the presence of such effects affects only inference, in the CIC model these can cause bias in the estimates. Fortunately, the presence of these and other kind of violations of the assumptions of the CIC model can be tested using the above result.

Since for each outcome I can base the specification test on either ATE or any of the QTE parameters, there is a large number of test results one could potentially present. Instead of showing all of these results, I report in table 5 only the proportions of ATE- and QTE-based specification tests that reject the null hypothesis at significance levels 0.05 and 0.10. Overall, the results from the specification test provide strong support for the validity of the CIC model for both test scores and earnings, and this is also reflected in the proportions of rejections reported in the table. These proportions are in general relatively small and always either close to or far below the given significance level, suggesting that the few rejections of the \( H_0 \) observed are consistent with being due to sampling variation.

One feature that separates the CIC model from the standard DID model is the asymmetric treatment of cohorts and regions in the CIC assumptions. However, it is possible to reverse the roles of cohorts and regions in the identification assumptions and estimate what Athey and Imbens (2006a) refer to as the reverse CIC model. Given a continuous outcome, these two models are empirically indistinguishable from each other and yield potentially different estimates for the
To check whether my results are sensitive to the way cohorts and regions are treated in the identification assumptions I have re-estimated the models using reverse CIC. In figures 10 and 11 I show the QTE estimates for the effect of early tracking on the quantiles of the test score and earnings distribution obtained using this alternative approach. While there are some small differences between the CIC and reverse CIC estimates, the overall picture one gets using the two alternative sets of identifying assumptions is practically identical. While in general there is no reason for this to be the case, it is nevertheless relieving to find that my results are not sensitive to such modification in the assumptions of the model.

As a final check on the validy of the identification strategy I have defined a placebo reform using only cohort-region cells that all attended either the tracked or the comprehensive school system. I begin by arbitrarily choosing 9 cohort-region cells to attend the tracked school system and then define baseline cohorts and control regions for them in such a way that there are no actual changes in the school system attendend among the treatment and control regions and the treatment and baseline cohorts. The QTE estimates for the effect of early tracking on the quantiles of the test score and earnings distributions using this placebo reform are shown in figures 12 and 13. Reassuringly, the estimated effects are statistically insignificant for all of the quantiles and also mostly small in magnitude.
5 Conclusions

There is substantial heterogeneity across countries in the extent to which students are tracked either within or between schools based on their ability. In this paper I shed light on the long-term distributional consequences of such tracking by studying the effects of a school system that divided students into an academic and non-academic track at a relatively early age. For this purpose I utilize an educational reform in Finland that replaced the old tracked school system with a comprehensive school system and postponed the tracking of students from age 11 to age 16. The comprehensive school reform was gradually implemented across the nation during the period 1972-1977, generating a promising quasi-experimental setting for capturing the causal effects of early tracking.

According to my results, early tracking had neither average nor distributional effects on the performance in a cognitive skills test taken by conscripts in the Army typically around age 20. However, I find that early tracking, on average, increased annual earnings around age 30-41. This effect seems to have come from a widening of the earnings distribution that was due to an increase in earnings especially in the highest quartile of the distribution. The bottom half of the earnings distribution was instead unaffected by early tracking. I find also substantial heterogeneity in the earnings effects according to parental background. It seems that the positive effects of early tracking were concentrated among individuals with educated or high earning parents. There seems to be instead no heterogeneity in the effects of early tracking on test scores along these dimensions.

One way of looking at these findings is by considering the trade-off between efficiency and equity which is often at the heart of the debate over tracking and selective school systems. The positive average effect found for earnings suggests that early tracking can have efficiency-increasing consequences for the economy. These consequences can, however, come at the cost of increasing inequality in the economy as illustrated by my result on the distributional effects of early tracking on earnings. On the other hand, it is interesting that this widening of the earnings distribution does not seem to have been accompanied by a reduction of earnings in the bottom half of the distribution.

Taken together, these results illustrate the important long-term consequences that education policy can have both on the individuals and the economy as a whole. While the focus of this paper has been on the timing of tracking student into an academic and non-academic track, my findings can also have potential relevance for various other policies that tend to increase the segregation of students according to ability and parental background. More research is needed for understanding better the possible distributional consequences of such policies.
References


Appendix

Figure 7: QTE estimates for the effect of early tracking on the quantiles of the test score distribution separately for each test section

Note: Dashed lines represent 95% confidence interval
Figure 8: QTE estimates for the effect of early tracking on the quantiles of the earnings distribution separately for years 1996-1999

Figure 9: QTE estimates for the effect of early tracking on the quantiles of the earnings distribution separately for years 2000-2003
Table 4: ATE estimates for the effect of early tracking on test scores and earnings for subsamples based on parental education and earnings

<table>
<thead>
<tr>
<th></th>
<th>DID</th>
<th>CIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Uneducated parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>-0.016</td>
<td>-0.014</td>
</tr>
<tr>
<td>(0.014)</td>
<td>(0.014)</td>
<td></td>
</tr>
<tr>
<td>Earnings</td>
<td>131</td>
<td>139</td>
</tr>
<tr>
<td>(196)</td>
<td>(200)</td>
<td></td>
</tr>
<tr>
<td><strong>Educated parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>0.007</td>
<td>0.011</td>
</tr>
<tr>
<td>(0.012)</td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td>Earnings</td>
<td>473**</td>
<td>455**</td>
</tr>
<tr>
<td>(202)</td>
<td>(212)</td>
<td></td>
</tr>
<tr>
<td><strong>Low earning parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>-0.013</td>
<td>-0.004</td>
</tr>
<tr>
<td>(0.014)</td>
<td>(0.016)</td>
<td></td>
</tr>
<tr>
<td>Earnings</td>
<td>69</td>
<td>35</td>
</tr>
<tr>
<td>(191)</td>
<td>(245)</td>
<td></td>
</tr>
<tr>
<td><strong>High earning parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>0.001</td>
<td>-0.009</td>
</tr>
<tr>
<td>(0.012)</td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td>Earnings</td>
<td>490**</td>
<td>579***</td>
</tr>
<tr>
<td>(206)</td>
<td>(208)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table reports DID and CIC estimates of ATE for subsamples based on parental education and earnings. Standard errors in parentheses. * significant at 10%, ** significant at 5%, *** significant at 1%.
Table 5: Proportion of ATE- and QTE-based specification tests rejecting the null hypothesis at a given significance level

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full sample:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>0.010</td>
<td>0.040</td>
</tr>
<tr>
<td>Earnings</td>
<td>0.010</td>
<td>0.020</td>
</tr>
<tr>
<td><strong>Uneducated parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>0.020</td>
<td>0.040</td>
</tr>
<tr>
<td>Earnings</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Educated parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>0.060</td>
<td>0.110</td>
</tr>
<tr>
<td>Earnings</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Low earning parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>0.050</td>
<td>0.100</td>
</tr>
<tr>
<td>Earnings</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>High earning parents:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test score</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Earnings</td>
<td>0.000</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Notes: This table reports the proportion of ATE- and QTE-based specification tests rejecting the null hypothesis at a given significance level. The proportions are reported for the full sample and subsamples based on parental education and earnings.
Figure 10: QTE estimates for the effect of early tracking on the quantiles of the test score distribution using reverse CIC

Figure 11: QTE estimates for the effect of early tracking on the quantiles of the earnings distribution using reverse CIC
Figure 12: QTE estimates for the effect of early tracking on the quantiles of the test score distribution using a placebo reform.

Figure 13: QTE estimates for the effect of early tracking on the quantiles of the earnings distribution using a placebo reform.