University of Zurich[UZH]
URPP Equality of Opportunity

And Tracking for All: Causes and Effects of Pupil Sorting in Middle School

Tuomas Kari

Equality of Opportunity Research Series #38
November 2023

URPP Equality of Opportunity

URPP Equality of Opportunity Discussion Paper Series No.38, November 2023

And Tracking for All: Causes and Effects of Pupil Sorting in Middle School

Tuomas Kari
University of Zurich /Aalto University
tuomas.kari@econ.uzh.ch

And Tracking for All: Causes and Effects of Pupil Sorting in Middle School

Tuomas Kari[a,b]

[a]University of Zurich
[b]Aalto University

27.11.2022

Abstract

Tracking, the policy of separating pupils into groups based on aptitude, is common, controversial and imperfectly understood. Little consensus exists on the circumstances under which tracking is practiced and what effects it may have on pupils. In this paper, I develop a novel method of measuring within-school tracking using observational data and estimate its long-run effects across a broad set of pupil outcomes. I show that tracking is prevalent, and that it varies both across schools and within schools over time. I find only limited evidence for tracking having significant short or long-run effects on pupils, although girls and boys seem to be affected differently. Finally, I provide evidence against the notion that tracking is a driver of inequality.

This work was supported by the University Research Priority Program "URPP Equality of Opportunity" of the University of Zurich.

1 Introduction

Systematic sorting of pupils to educational groups, structured hierarchically based on performance or perceived ability, a practice also known as tracking, is a matter of contention among policy makers and educational researchers. Governments, school administrations, and parents tend to have different and sometimes contradictory considerations when forming educational groups. Some aim to provide equal opportunities to all pupils, others prefer to promote elite outcomes to educate the next generation of leaders. By consequence. even assuming that the effects of tracking were well-understood, optimal tracking policy would still depend on the objective function of the policy makers and may vary accordingly. This paper seeks to identify and understand the circumstances under which schools elect to engage in tracking, and then evaluate the policy's long-run effects on pupil outcomes.

The fact that official tracking practices are generally uniform within countries makes studying the topic empirically challenging. While the sorting of pupils into groups can take place at different organizational levels - such as into separate schools, into special study programs, or simply into normal classes within schools - the policy is set at an aggregated level. Consequently, there is little granular geographical variation in tracking that could be utilized for statistical analysis; comparisons at a higher level of aggregation, such as across countries, comparisons have their own obvious issues.

In this paper, I develop a novel method of measuring tracking using observational data. This allows one to compare schools with varying levels of tracking while holding other institutional characteristics constant. The measure of tracking intensity is based on how concentrated seemingly high-ability pupils, proxied by parental educational background, are across classes for each cohort entering a school. Conceptually, tracking intensity can be estimated by comparing the observed distribution of pupils with college-educated parents across classes to a simulated, random distribution. The more the observed distribution differs from the simulated distribution, the higher the likelihood that the allocation process is non-random, and thus the higher the intensity of tracking. I utilize the nonparametric Kolmogorov–Smirnov test to quantify this distance between the empirical distribution function of the sample and the simulated distribution function.

Applying the measure to pupils in Finnish middle schools in 1996-2011, I show that tracking is prevalent, and that it varies both across schools and within schools over time. In the first part of the paper, I seek to understand what determines a school's chosen level of tracking. I provide descriptive evidence of institutional responses to changes in pupil composition. By studying variation within schools over time and controlling for national level time trends, I show that a higher share of Finnish-speaking pupils and of pupils with better-educated parents in a school-cohort are correlated with higher tracking. Higher shares of first or second generation immigrant pupils of Middle Eastern or African background are associated with lower intensity of tracking. I study whether the level of urbanization affects the descriptive results and find more muted responses to changes in pupil composition in cities. In fact, the result that parental education is associated with higher tracking is fully driven by non-urban areas.

Due to concern that changes in pupil composition might be endogenous to other determinants of tracking intensity, I use a Bartik instrument to verify that the share of immigrants causally affects tracking intensity. The results are inconclusive, however, and I find no robust relationship between ethnic composition and tracking intensity.

In the second part of this paper, I shift the attention to the long-run effects tracking has on pupils. I rely on the two-way fixed effects, that utilize over-time variation within schools, for causal identification. After showing that, at the individual level, tracking is not associated with pupil and parental background variables, I study a set of long-run outcomes: middle school grade-point average, graduating from an upper secondary and college, criminality, unemployment and earned income.

On average, tracking does not have statistically significant effects on pupil outcomes, except perhaps a negative one on the likelihood of college graduation. However, I provide evidence that pupils are differentially affected by tracking depending on their gender. More intense tracking seems to be detrimental for girls and possibly of limited benefit for boys. Female pupils earn less as adults and are somewhat less likely to graduate from college when tracking increases. For male pupils, more intense tracking increases their likelihood of completing upper secondary schooling. Separating the effects of tracking by the deciles of parental income does not yield any clear patterns; if anything, tracking benefits mostly pupils in lower-income groups. Higher tracking does not seem to increase inequality of pupil outcomes. Conditional on the two-way fixed effects, tracking is not significantly associated with the standard deviations of the set of outcomes. I also find no evidence for tracking decreasing intergenerational mobility, measured as the correlation between parental and pupil income.

Overall, I find that tracking, i.e. the systematic sorting of pupils into classes within a school cohort is common and its intensity varies across schools and over time. Interestingly, tracking intensity seems to respond to changes in the composition of the student body. On the other hand, I find only limited evidence for tracking having significant short or long-run effects on pupils. I can rule out the hypothesis that it drives inequality, both in absolute terms and relative to parental income. What sets this analysis apart from previous literature is that it evaluates the implications of tracking for the entire population of pupils broadly and not just at some margin defined by an institutional reform or a cutoff in test scores. This is made possible by the novel method of measuring tracking intensity in observational data. Furthermore, the richness of the Finnish administrative data allows me to study many different outcomes over the pupils' post-school careers and, with links to family members, to estimate effects on inter-generational mobility.

Tracking has been traditionally rationalized as an efficient way to prepare pupils to the different career paths to which they are naturally best suited. Pupils were seen as differing fundamentally from one another in ability, motivation and aspirations. By this line of thinking, tracking would fulfill two important requirements; it would service pupils by fitting academic demands better to the level of individual pupils, and it would help produce workers with a variety of skill-sets to meet the demands of the economy. In recent decades, tracking has been discussed mostly in terms of the equity of pupil outcomes. In many developed countries, school systems were reformed to reduce tracking in the second half of the 20th century. In the UK and the Nordic countries, for example, this took the form of implementing comprehensive school systems. Nevertheless, the practices of tracking vary widely across countries and school systems. There are significant differences in at what age pupils are separated into tracks and in whether tracking takes place at all.

From a theoretical viewpoint, the effects of tracking on achievement are uncertain a priori. On the one hand, one could expect that the reduced heterogeneity within educational groups might make teaching easier. Teachers would not have to adjust the material and the level of demands to different pupils within the same class, benefiting all pupils Lazear (2001).

On the other hand, tracking creates differences in peer composition. There is an extensive literature on peer effect, which indicates that changing the set of peers that a pupil is exposed to can have significant effects on learning outcomes. Tracking might increase inequality through peer assignment; if pupils generally benefit from talented peers, tracking should help stronger pupils that get matched with equally strong peers and hurt weaker ones placed in a lower-level group (Epple et al., 2002)

In many practical cases, tracking induces differences also in curricula, the material taught to the pupils, and teacher quality. These changes are likely to affect the pupils, but in less predictable ways compared to the aforementioned mechanisms. While tracking may allow teachers to adapt their teaching more accurately to pupil level, which may improve their understanding and self-esteem, teachers might also sub-optimally lower their expectations for the pupils in low- and middle-ability classes and consequently teach them lower-level skills and knowledge, and expose them to less diverse instructional materials (Peltier, 1991).

There is an extensive empirical literature studying the effects of tracking. The variation in the institutional settings present and methodological approaches employed in that literature mean that the results have not always been directly comparable and few set-in-stone conclusions have emerged.

Perhaps the clearest evidence comes from an experiment in Kenya where Duflo et al. (2011) randomly selected some schools to conduct tracking across classes and showed that it benefits everyone. In developed countries, researchers have used multiple methods to study the effects of tracking but to the best of my knowledge, similar experimental evidence is lacking. Dustmann et al. (2016) use an instrumental variable approach to estimate the effects of a pupil being allocated to academically-focused Gymnasium. For the marginal pupils, attending Gymnasium has no long-run effects.

Many papers take advantage of school system reforms that change the intensity of tracking or the age at which it happens. Typically, reforms reducing tracking also reduce inequality in pupil achievement, and in some cases may improve achievement on average (Piopiunik (2014); Pekkala Kerr et al. (2013); Canaan (2020)). Cross-country evidence, where causal identification is less credible, too is mixed (Hanushek and Woessmann (2011); Brunello and Checchi (2007)). Evaluating one special program, Card and Giuliano (2016) show that allocating talented pupils to a separate class within schools in the US benefits high-achieving minority pupils.

Evidence of within-school tracking in developed countries has been limited until late. To the extent that reliable estimates existed, the identification strategies employed mean that the effect is estimated for pupils on the margin. Therefore, we lacked a firm understanding of how run-of-the-mill tracking happens within schools and how the bulk of the pupil population is affected. In a recent contribution to the literature, Antonovics et al. (2022) find substantial within-school tracking in the Texas public school system.

The diversity in tracking policies across countries, school systems, and over time points to significant disagreements on the desirability of such policies. Educational systems are shaped by historical contingencies and each country's idiosyncratic circumstances. Nonetheless, one must assume educational authorities are able to exert some influence over tracking intensity. The choice of tracking intensity is likely to reflect pupil composition. As the set of pupils entering a school changes, policymakers and parents will adapt their desired level of tracking, both for their own reasons. Policymakers might have the best interest of all pupils in mind but they are also subjected to pressure from the parents who are more focused on the well-being of their own children. If pupil heterogeneity increases, peer effects might become more pronounced and parents of presumed high-ability children should increase their efforts to lobby policymakers to instate tracking policies. The ability of parents to influence policy will also likely vary, perhaps depending on their educational background or socioeconomic status.

The aforementioned study by Antonovics et al. (2022) provides the best empirical exploration into the determinants of tracking to date. In their setting, tracking is primarily driven by heterogeneity in the pupil population. Their work notwithstanding, the main challenge in the literature has been that the variation in tracking takes place mostly across countries or over time at the national level. A more granular measure of tracking is necessary for properly analysing which factors determine the level of tracking chosen by schools and/or municipalities. I provide a data-driven, alternative approach to measuring tracking, which does not require as detailed information on the prior ability of the pupils as the one used by Antonovics et al. (2022).

The rest of the paper proceeds as follows. In the next section, I describe the Finnish school system and the data. In the third section, I introduce the measure of tracking intensity based on the Kolmogorov-Smirnov test. Then, I present the results of an analysis on the determinants of tracking intensity. In the fifth section, I move to the second main approach of the paper and study the effects of tracking on the pupils. The sixth section focuses on the effects on inequality, and the seventh concludes.

2 Setting and data

2.1 Institutional setting

The Finnish school system is based around the comprehensive school, which lasts 9 years from age 7 to age 16. All pupils follow the same curriculum in the comprehensive school; that is, there is no explicit tracking until age 16. At this point, pupils apply to institutions of secondary education through a centralized system. The secondary institutions are split into academic upper secondary schools and vocation schools. Admission is based on middle school grades.

The 9-year-long comprehensive school is *de facto* split into two parts. Pupils first attend a primary school for the first 6 years and then move to middle schools with new class allocations. This juncture is highlighted in red in Figure 1 that illustrates the school system.

Policy discussion in Finland as well as academic research have paid considerable attention to the growing skills gaps between schools and differences in demographic structure of the school catchment areas (Bernelius (2013), Vilkama and Bernelius (2019)). However, the allocation of pupils to classes within schools, and the possible inequalities created by so doing, is less explored, except in a case-study on a single city (Berisha and Seppänen, 2017).

There are few national rules governing the class allocation. Municipalities are responsible for educational provision and are free to decide on class allocation as long as it is consistent with pupils' right to adequate education. In practice, the decision-making is typically relegated to middle school principals. They consult primary school teachers familiar with the pupils and take into account pupils' wishes on friends with whom they would like to be placed into the same class. In some middle schools, pupils or their parents can also apply to special classes that focus on a specific subjects such as a foreign language, music, or science.

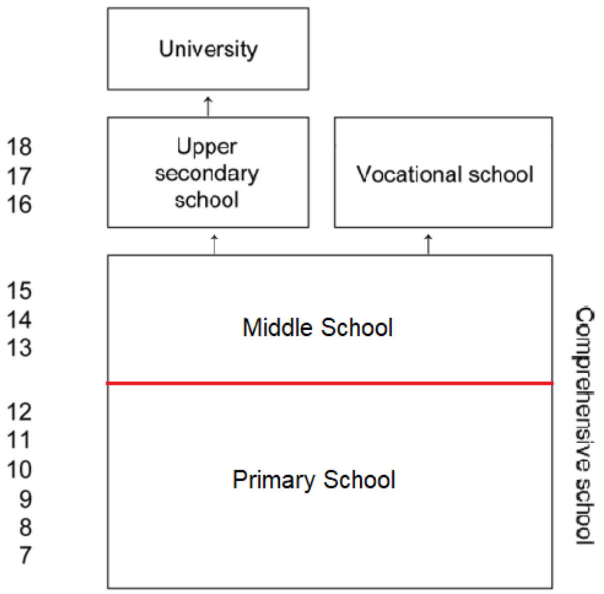

Figure 1: The Finnish school system

There are multiple aims for the class allocation. According to school principals that I have consulted, the focus in class allocation is on maintaining social connections from primary school and creating a harmonious learning environment. Primary school grades are typically not considered directly and some municipalities instruct schools to avoid separating pupils based on academic ability. Principals always take into account any needs for special, in-class support and try to break up known bully-bullied pairs.

Given the leeway that school principals and administrators enjoy in deciding on the class allocation, there is likely to be substantial variation in how systematically class allocations reflect pupil ability or social background.

2.2 Data description

The research utilizes individual-level panel data created by Statistics Finland by linking information from several administrative registers. For creating the index of tracking intensity, I utilize data on middle school class allocations, which come from the centralized secondary education application system and covers the years 1996 to 2011. The individuals applying in those years are born between 1980 and 1995. The data include information for all applicants on their school, cohort, class, middle school and an individual identifier based on social security numbers.

Using the identifier, I combine the data with information on long-run outcomes originating in other registers. These include yearly labor market outcomes including earned income and unemployment, secondary and tertiary education and criminality. I analyze pupil background using information on their gender, birth country, immigrant-background and mother tongue. I link the pupil information to their social parents, that is adults living in the same household, for whom I observe yearly labor market outcomes such as income and unemployment, education, gender, birth country, immigrant-background, and mother tongue.

3 Index of Tracking Intensity

There is no generally accepted method of measuring the intensity of tracking within schools across seemingly similar classes. Conceptually, a tracking measure should capture how systematic the allocation of pupils into classes is within a school cohort. A random allocation should yield a low value of tracking and make pupil background similar across classes. A highly sorted distribution should lead to a high value for the tracking measure and to pupils in one class differing significantly from those in another.

In order to operationalize the concept, one must first choose on which dimension to compare pupils. Typically tracking is done to place pupils of similar ability to the same groups. A uni-dimensional measure of latent ability would therefore be the ideal pupil characteristic on which to base a tracking index, but such a thing is unfortunately not observable in the data available. I utilize a binary variable capturing whether any of the pupil's social parents, i.e. adults living in the same household, have completed a college-level degree to proxy for pupil ability.

The next step is to choose the method of measuring how systematic the allocation of pupils is. My preferred measure of tracking intensity is the two-sample Kolmogorov–Smirnov test (Massey, 1951). In essence, the test measures the likelihood that the empirical distribution of pupils in classes and a simulated, random allocation could have been drawn from the same probability distribution. More precisely, it measures the difference between the two distributions as the largest absolute difference between the two cumulative distribution functions.

$$\text{Kolmogorov-Smirnov divergence: } D_{KS} = max_i |h_1(i) - h_2(i)|$$

The Kolmogorov-Smirnov test is attractive, because it is nonparametric and picks up differences in the shape as well as in level shifts in the cumulative distribution functions. The metric can also be illustrated graphically as the largest vertical distance between the two CDF curves and takes values between 0 and 1.

I create the Kolmogorov-Smirnov measure at the year-school level, given that school principals have final say on class allocations and the decisions are made separately for each cohort. There are 10079 school-cohort cells. Looking at pooled data, the tracking index varies substantially, with mean 0.35 and standard deviation 0.12. Figure 2 shows the histogram of the tracking index.

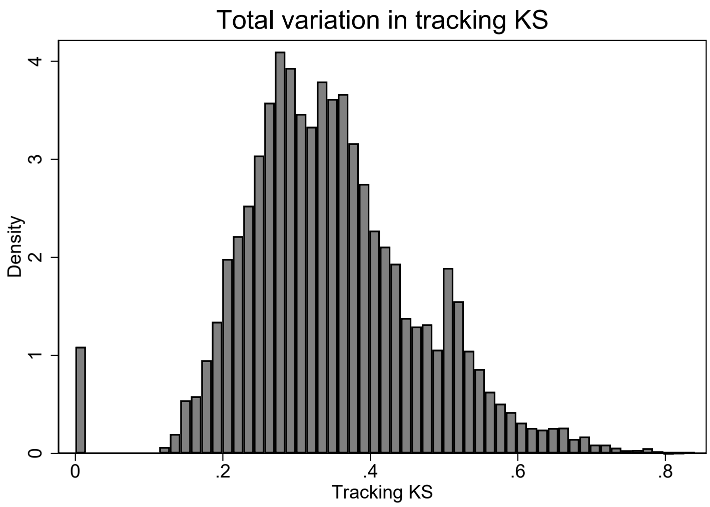

Figure 2: Pooled variation in tracking intensity

Around one third of the variation in tracking takes place within schools, over time, as can be seen in Figure 3, which depicts the histogram of standard deviations computed separately for each school. The mean standard deviation is 0.10. The within-school variation is crucial for my empirical strategy, as we will see in the section.

Relating the within-school standard deviation to the full-sample mean of 0.35 provides another perspective to the variation of the tracking measure. On average, within a school tracking varies by 28 percent of the full-sample mean. The standard deviation of this measure is a further 9 percentage point. There is, therefore, substantial variation in tracking intensity over time.

Figure 4 illustrates the geographical variation in pooled data collapsed to the municipality-level. It is immediately clear from the map that the data does not cover all the municipalities in the country. This is largely due to deficiencies in the way some schools have reported the information on class allocation. More encouragingly, the variation of the tracking intensity doesn't seem to follow any obvious geographical pattern. One could argue that tracking is lower in the eastern part of the country, but even there some municipalities track at a nationally high level.

Figure 5 shows that the national average of the tracking intensity has remained remarkably stable over the time period under study. There is no discernible trend visible, even though the same period has witnessed secular changes in society and in pupil composition as the size of the foreign-born population more than doubled.

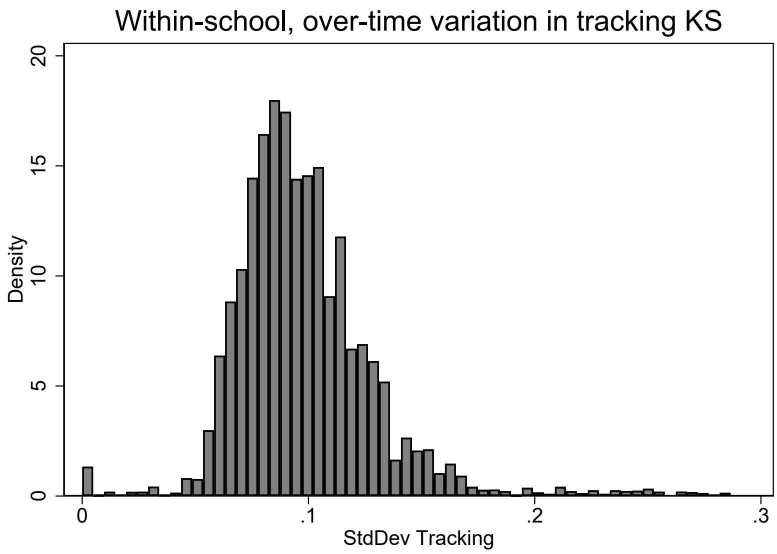

Figure 3: Within-school variation in tracking intensity

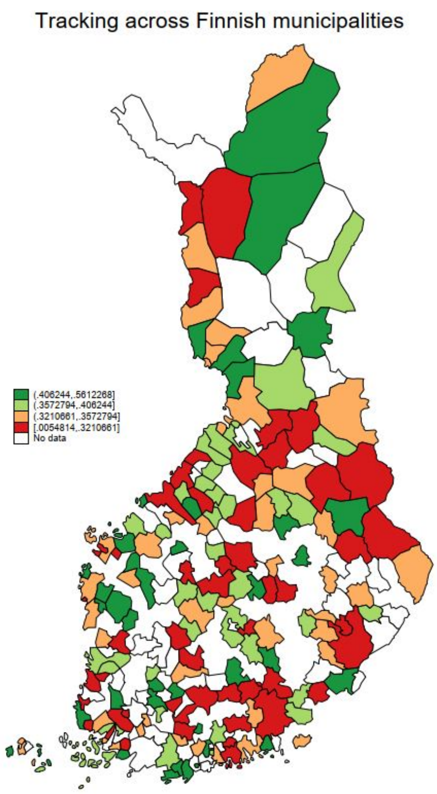

Figure 4: Geographical variation in tracking intensity

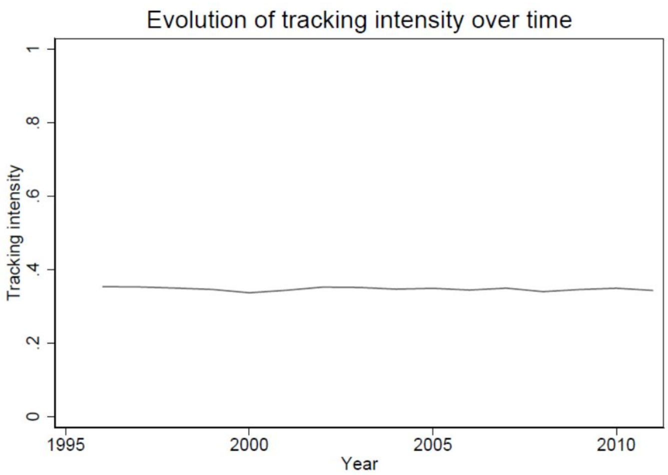

Figure 5: Evolution of tracking intensity

4 Determinants of Tracking

In the first application of the tracking index, I will investigate the determinants of tracking intensity, i.e. which factors are associated with tracking intensity, conditional on cohort and school fixed effects. Formally, I want to estimate the β_1 in the following equation:

$$T_{sc} = \beta_0 + \beta_1 X_{sc} + \alpha_c + \alpha_s + \epsilon_{sc}$$

where s denotes school, and c denotes cohort. T_{sc} is the index for tracking intensity, X_{sc} is a vector of variables describing the pupil composition at school s for cohort c, a_c is a cohort fixed effect, a_s is a school fixed effect.

The two-way fixed-effect setup eliminates national trends and time-invariant differences between schools. If there are no other factors that affect tracking intensity and change together with the pupil composition, we can interpret β_1 as capturing the causal effect of the relevant dimension of pupil composition on tracking intensity. I do not want to overstate this assumption, though, and the results should primarily be seen as descriptive.

It becomes clear that changes in the pupil composition are systematically associated with tracking intensity. In Table 1, we can see that higher parental education is associated with higher tracking intensity. This is true regardless of which level of educational achievement one studies. Interestingly, parental income is not associated with the tracking intensity.

Table 1: School-level determinants of tracking: How parental education and income are associated with tracking intensity

	(1) Tracking KS	(2) Tracking KS	(3) Tracking KS	(4) Tracking KS
Share with upper secondary schooled parents	0.107^{***} (0.0245)			
Share with college educated parents		0.0800^{***} (0.0275)		
Share with graduate schooled parents			0.0972^{***} (0.0337)	
Parental mean income at 50, €k				-0.000169 (0.000473)
Year FE	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes
N	10078	10078	10078	10078
Y mean	0.353	0.353	0.353	0.353

Standard errors in parentheses
School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2 shows that the ethnic background of pupils also plays a role. Having a higher share of pupils who are native Finnish-speakers is associated with higher tracking. The other notable result is that tracking is lower when the share of pupils born in or with parents born in the Middle East or Africa is higher. The same pattern does not apply to foreign pupils from Western countries or Eastern Europe. In facts, when looking at the birth place of the parents, a higher share of pupils with foreign, Western heritage is associated with higher tracking.

Table 2: School-level determinants of tracking: How ethnic composition of pupils is associated with tracking intensity

	(1) Tracking	(2) Tracking	(3) Tracking	(4) Tracking	(5) Tracking	(6) Tracking	(7) Tracking	(8) Tracking	(9) Tracking
Share immigrant pupils	-0.00619								
	(0.0547)								
Share Finland-born pupils		0.0263	-0.0124	0.0428					
		(0.0514)	(0.0592)	(0.0520)					
Share foreign, Western-born pupils		-0.278							
		(0.327)							
Share ME-Africa-born pupils			-0.264**						
			(0.123)						
Share East Europe-born				0.217					
				(0.167)					
Share with immigrant parents					0.0636				
					(0.0458)				
Share only Finland-born parents						-0.0125	-0.105	0.00925	
						(0.0517)	(0.0692)	(0.0530)	
Share foreign, Western-born parents						0.252***			
						(0.0331)			
Share ME-Africa-born parents							-0.383***		
							(0.115)		
Share East Europe-born parents								-0.290	
								(0.412)	
Share Finnish speaking									0.201***
									(0.0298)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	10078	10078	10078	10078	10078	10078	10078	10078	10078
Y mean	0.353	0.353	0.353	0.353	0.353	0.353	0.353	0.353	0.353

Standard errors in parentheses

School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As one can see in Table 3, the level of urbanization of the school catchment area modulates the relationship between tracking intensity and pupil composition. When none of the school's pupils live in what the Finnish Environment Institute classifies as an urban location, a higher share of pupils with college-educated parents is associated with higher tracking intensity. In more urban schools, parental education is actually associated with lower tracking. The dynamic that determines tracking intensity clearly varies across locations. In places where the the share of pupils living in urban areas is higher, higher parental income seems push down tracking intensity. There's some evidence that in urban areas a higher share of native-born pupils drives tracking intensity up.

The observed association between these factors and tracking intensity does not allow us to conclude that they have a causal effect on tracking. Plausible identification is luckily available for at least the immigrant shares. I use an instrumental variable approach to estimate the causal effect of changes in the share of immigrants on tracking. As is typical in the immigration literature (**?**), I utilize a shift-share or Bartik instrument (**?**). I predict the change in immigrant share in a school area by interacting the lag of the school-level immigrant shares with national growth rates in the relevant population. The results are shown in Table 4, but none of them are statistically significant. This might reflect an authentic lack of a causal relationship or the decrease in statistical power. The instrumental variable regressions can only use part of the total variation in ethnic composition at the school level. The uncertainty means that the additional causal identification has not changed the picture.

Table 3: School-level determinants of tracking: How pupil composition is associated with tracking intensity

	(1) Tracking	(2) Tracking	(3) Tracking	(4) Tracking	(5) Tracking	(6) Tracking
Share in city	-0.0286***	-0.00926	-0.224*	-0.0318	0.0314	0.0376
	(0.00325)	(0.0334)	(0.123)	(0.147)	(0.0357)	(0.0398)
City X Share immigrant pupils		-0.278				
		(0.184)				
Share immigrant pupils		0.220				
		(0.166)				
City X Share Finland-born			0.218*			
			(0.123)			
Share Finland-born pupils			-0.104			
			(0.0972)			
City X Parent born in ME-Afr				-0.335		
				(0.369)		
Share ME-Africa-born parents				-0.0945		
				(0.311)		
City X All parents Finland-born				0.0182		
				(0.146)		
Share only Finland-born parents				-0.118		
				(0.114)		
City X College educated parent					-0.187***	
					(0.0510)	
Share with college educated parents					0.173***	
					(0.0388)	
City X Parent income						-0.00168**
						(0.000740)
Parental mean income at 50, k						0.00115
						(0.000770)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	No	Yes	Yes	Yes	Yes	Yes
N	10078	10078	10078	10078	10078	10078
Y mean	0.353	0.353	0.353	0.353	0.353	0.353

Standard errors in parentheses

School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Determinants of tracking: Results from Shift-share instrumentation

	(1) Tracking	(2) Tracking	(3) Tracking
Share immigrant pupils	1.543		
	(1.06)		
Share ME-Africa-born pupils		0.813	
		(0.29)	
Share foreign, West-born pupils			72.98
			(0.16)
Share Finland-born pupils		0.142	1.799
		(0.29)	(0.16)
Year FE	Yes	Yes	Yes
School FE	Yes	Yes	Yes
N	9108	9108	9108
Y mean	0.353	0.353	0.353

t statistics in parentheses

School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5 Effects of Tracking

5.1 Empirical strategy

In the second application of the tracking index, my goal is to estimate the causal effects of the intensity of tracking on the long-run outcomes for pupils. The causal identification is based on two-way fixed effects, meaning that cohort fixed effects control for national-level time trends and school fixed effects control for time-invariant differences across school catchment areas. The remaining variation reflects changes in tracking intensity over time.

$$Y_{isc} = \beta_0 + \beta_1 T_{sc} + \beta_2 X_{isc} + \alpha_c + \alpha_s + \epsilon_{isc}$$

where i denotes individual, s denotes school, and c denotes cohort Y_{isc} are the outcomes: middle school GPA, upper secondary school graduation by 25, college graduation by 25, criminality by 25, earnings at 30, unemployment at 30, T_{sc} is the standardized index for tracking intensity, X_{isc} is a vector of individual control variables, a_c is a cohort fixed effect, a_s is a school fixed effect

5.2 Balancing tests

The causal analysis relies on the identifying assumption that, conditional on the national-level time trends, the changes in tracking intensity for a school are not correlated with other changes that affect pupil outcomes. In other words, no other factors that drive pupil outcomes change in conjunction with tracking intensity. To assess the plausibility of the assumption, I test whether student background is balanced across different values of tracking intensity. Table 5 shows the results from regressing tracking intensity on a host of individual characteristic, and year and school fixed effects. None of the pupil characteristics is significant at the 5 percent level, and only one, having parents that graduated from upper secondary school breaks the significant-at-10-percent-level line. However, in the previous section we already showed that tracking intensity is driven by changes in the pupil composition. Perhaps the fact that one cannot detect an association with tracking intensity and pupil background at the individual-level is moot and the following analysis is fundamentally invalid. That notwithstanding, the next subsection presents the results.

5.3 Results

The first step is to estimate the average effect of tracking intensity on a set of long-run, pupil outcomes: middle school GPA, graduation from upper secondary school by age 25, graduation from college by age 25, having a criminal record by age 25, months spent unemployed at age 30, and earned income at age 30. Table 6 shows that on average tracking does not affect pupil outcomes.

The lack of discernible effects on average, might hide the way tracking affects sub-populations. If tracking hurts one group and helps another, one would not necessarily detect any effect in the pooled sample. My results show that gender might play just such a role. Tables 7 and 8 show the effects of tracking when the sample is limited to female and male pupils respectively. For women, higher tracking reduces the likelihood of having a college degree and earnings at age 30. The effect on college graduation is minimal and only significant at the 10 percent level, but for income tracking plays a non-trivial role. A standard deviation change in tracking intensity reduces yearly income by 95 euros. This accounts for just under half a percent of the average female yearly earnings. For men, tracking is less consequential. Tracking intensity seems to improve male pupils' chances of graduating from upper

Table 5: Balance: How parent and pupil background characteristics are associated with tracking intensity. Results from separate regressions.

	(1) Tracking
Any parent graduated upper secondary school	0.000383*
	(0.000231)
	[894448]
Any parent graduated college	0.0000913
	(0.000256)
	[839696]
Any parent graduated Masters or PhD	0.000366
	(0.000313)
	[839696]
Any parent immigrant	-0.000347
	(0.000651)
	[895520]
All parents born in Finland	0.000959
	(0.000666)
	[895520]
Pupil of immigrant background	-0.00101
	(0.00101)
	[891619]
Pupil female	-0.00000743
	(0.000225)
	[891619]
Pupil Finnish-speaking	-0.000366
	(0.00104)
	[895504]
Parental mean income at 50	-4.04e-09
	(5.07e-09)
	[893574]
Year FE	Yes
School FE	Yes
Y mean	0.348

Standard errors clustered at school-year level and presented in parentheses. N in square brackets.
School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: How tracking intensity affects long-run pupil outcomes

	(1) Middle School GPA	(2) Upper secondary school	(3) College	(4) Crime	(5) Unemployment	(6) Earned income
Tracking	0.000953	0.000838	-0.000614	-0.000114	-0.00918	-20.41
	(0.00147)	(0.000707)	(0.000562)	(0.000344)	(0.0130)	(30.20)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
N	805536	895520	895520	895520	110351	514605
Y mean	7.758	0.474	0.218	0.074	4.838	28851

Standard errors clustered at school-cohort level, presented in parentheses.
School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

secondary school, but the effect is marginal both in statistical and economic significance. Even a conservative interpretation of these results allows us to conclude that tracking affects male and female pupils differently.

Table 7: The effects of standardized tracking intensity on female pupils

	(1) Middle School GPA	(2) Upper secondary grad	(3) College grad	(4) Crime	(5) Months unemployed	(6) Earned income
Tracking	0.000950	0.000115	-0.00160*	0.000000900	-0.00525	-95.26***
	(0.00192)	(0.000893)	(0.000825)	(0.0000103)	(0.0168)	(32.03)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
N	391928	444044	444044	444044	53820	254972
Y mean	8.047	0.552	0.291	0.000	4.382	25125

Standard errors clustered at school-year level, presented in parentheses.
School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: The effects of standardized tracking intensity on male pupils

	(1) Middle School GPA	(2) Upper secondary grad	(3) College grad	(4) Crime	(5) Months unemployed	(6) Earned income
Tracking	0.00120	0.00160*	0.000401	-0.000249	-0.0112	54.22
	(0.00204)	(0.000966)	(0.000653)	(0.000622)	(0.0189)	(47.81)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	No	No	No
N	410508	447572	447572	447572	56495	259623
Y mean	7.483	0.401	0.148	0.148	5.272	32511

Standard errors clustered at school-year level, presented in parentheses.
School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Similar analyses splitting the sample based on having immigrant parents, parents holding a college degree, or by parental income yield no statistically significant findings.

6 Effects on inequality

Previous theoretical and empirical research suggests that tracking could lead increased inequality, that is to a higher variation in pupil outcomes. I test this in two different ways, directly and from the perspective of inter-generational mobility. Table 9 shows the results from regressing the standard deviation of each outcome directly on tracking intensity. There is some evidence that more intense tracking could cause higher variation in educational outcomes. Column 3 shows a positive coefficient from the regression of years of schooling by age 30 on tracking intensity, but it is only significant at the 10 percent level. Tracking intensity does not seem to have a strong effect on the variation in pupil outcomes.

Table 9: The effects of tracking intensity on the standard deviation of outcomes

	(1) Middle school GPA	(2) Years of schooling by 25	(3) Years of schooling by 30	(4) Months unemployed	(5) Earnings at 30
Tracking	0.000731	0.00167	0.00741^{*}	0.00197	120.6
	(0.00111)	(0.00248)	(0.00391)	(0.0123)	(84.70)
Year FE	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes
N	10060	9468	6500	6826	6564
Y mean	0.866	1.430	1.833	3.652	15888

Standard errors in parentheses

School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.

Standard deviations in outcomes calculated within school and cohort.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Tracking, especially if it is based on parental characteristics, has the potential to further advantage pupils who might already have the most opportunities for self-advancement available to them. It can make it even more difficult for less-privileged children to compete in the educational or the economic arena. One could therefore hypothesize that tracking could reduce inter-generational mobility. Table 10 tests this by regressing pupil's earned income at age 30 on the interaction between above median tracking intensity and parental income. There is no evidence that tracking reduces inter-generational mobility. Figure 6 confirms this result in graphical form. There is no visible difference in the relationship between pupil and parental income between pupils from below and above median tracking intensity.

Table 10: How tracking intensity affects the correlation between parental and child income

	(1) Earned income at 30	(2) Earned income at 30	(3) Earned income at 30	(4) Earned income at 30
Parental mean income at 50, 1000	75.50^{***}	74.88^{***}	64.95^{***}	47.57^{***}
	(8.214)	(11.70)	(10.71)	(8.323)
Above median tracking		-160.5	-122.7	-80.29
		(502.9)	(452.9)	(348.2)
Above median tracking X parental income		1.232	2.112	1.015
		(16.37)	(14.68)	(10.92)
Year FE	No	No	Yes	Yes
School FE	No	No	Yes	Yes
Controls	No	No	No	Yes
N	513134	513134	513134	472715
Y mean	28872	28872	28872	29091

Standard errors clustered at school-year level and presented in parentheses.

School-year-level tracking created with Kolmogorov-Smirnov test with parental college graduation.

Parental controls: academic high school, college, MSc/PhD, birth country, immigrant background.

School and pupil controls: pupil born in Finland, mothertongue Finnish, immigrant , gender, number of parallel classes, number of students in class.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

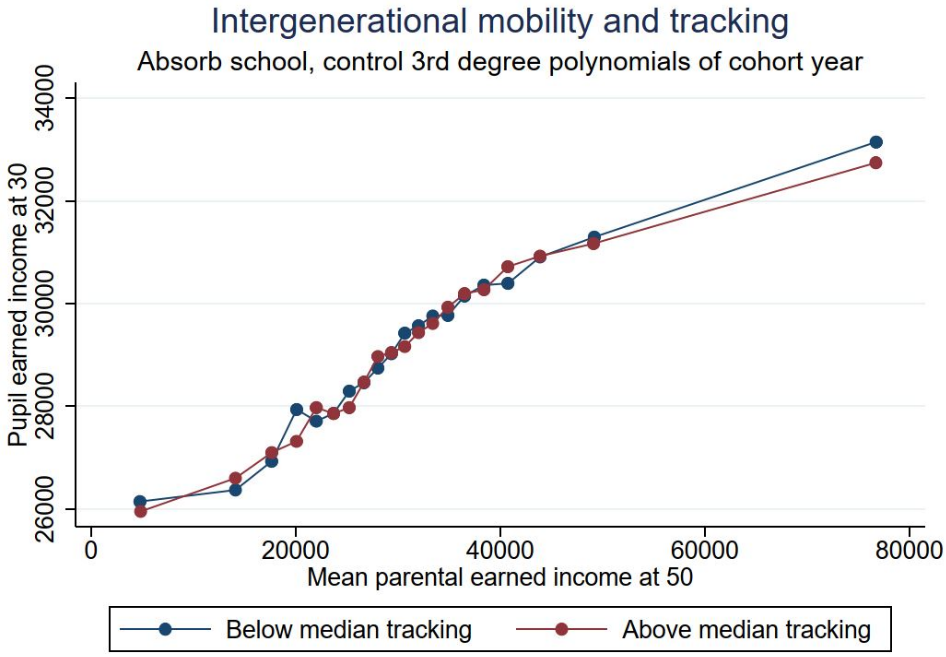

Figure 6: Effects of tracking on inequality: Intergenerational mobility

7 Discussion

This study creates a novel index that measures the intensity with which pupils are tracked across normal classes, that is, how systematically they are allocated to classes. With sufficiently rich observational data, the measure can be applied to may different settings. By measuring tracking intensity at the school-cohort level, I demonstrate that tracking intensity varies both across schools and over time.

My results show that schools react to changes in pupil composition by varying the intensity of tracking. I first document the various changes in pupil composition that are associated with changes in tracking intensity with schools. Parental education and pupils from culturally distant backgrounds stand out as important factors. I further show that school-specific factors, namely whether the school is located in an urban area, can have a crucial effect on how tracking is determined.

I then proceed to show that, while middle schools often engage in some level of tracking, this does not have significant results on the average pupil. However, this null result hides a more complex story. Intense tracking has negative effects on female pupils; they´re less likely to graduate from college and earn less when tracking is more intense. I find suggestive evidence that male pupils might on the other hand benefit from higher tracking intensity. the likelihood of graduating from upper secondary school is higher for male pupils when tracking is more intense. Contrary to previous literature, I find only marginally significant effects on inequality of outcomes. The results do not fully support the hypothesis that tracking would increase inequality.

In conclusion, my results suggest that the kind of tracking studied in this paper, i.e. systematically sorting pupils to classes within a middle school, is prevalent and responsive to changes in pupil composition. It is, however, perhaps not intense enough to produce the kinds of effects detected in other settings. I believe this research opens up new avenues in studying school tracking and it highlights the way in which tracking intensity is a flexible and potentially important policy variable.

References

ANTONOVICS, K., S. E. BLACK, J. B. CULLEN, AND A. Y. MEISELMAN (2022): "Patterns, Determinants and Consequences of Ability Tracking: Evidence from Texas Public Schools," *NBER Working Paper*.

BERISHA, A.-K. AND P. SEPPÄNEN (2017): " Pupil selection segments urban comprehensive schooling in Finland: composition of school classes in pupils' school performance, gender, and ethnicity," *Scandinavian Journal of Educational Research*, 2, 240–254.

BERNELIUS, V. (2013): "Eriytyvät kaupunkikoulut - Helsingin peruskoulujen oppilaspohjan erot, perheiden kouluvalinnat ja oppimistuloksiin liittyvät aluevaikutukset osana kaupungin eriytymiskehitystä," *City of Helsinki Research Series*.

BRUNELLO, G. AND D. CHECCHI (2007): "Does school tracking affect equality of opportunity? New international evidence," *Economic Policy*, 22, 782–861.

CANAAN, S. (2020): "The long-run effects of reducing early school tracking ," *Journal of Public Economics*, 187.

CARD, B. D. AND L. GIULIANO (2016): "Can Tracking Raise the Test Scores of High-Ability Minority Students?" *American Economic Review*, 106, 2783–2816.

DUFLO, B. E., P. DUPAS, AND M. KREMER (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101, 1739–1774.

DUSTMANN, C., P. A. PUHANI, AND U. SCHOENBERG (2016): "The Long-Term Effects of Early Track Choice," *Economic Journal*, 127, 1348–1380.

EPPLE, D., E. NEWLON, AND R. ROMANO (2002): "Ability tracking, school competition, and the distribution of educational benefits," *Journal of Public Economics*, 83, 1–48.

HANUSHEK, E. A. AND L. WOESSMANN (2011): *Chapter 2 - The Economics of International Differences in Educational Achievement*, vol. 3, Elsevier B.V., 1 ed.

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

MASSEY, F. J. (1951): "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, 46, 68–78.

PEKKALA KERR, S., T. PEKKARINEN, AND R. UUSITALO (2013): "School Tracking and Development of Cognitive Skills," *Journal of Labor Economics*, 31.

PELTIER, G. L. (1991): "Why Do Secondary Schools Continue to Track Students?" *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 64.

PIOPIUNIK, M. (2014): "Economics of Education Review The effects of early tracking on student performance : Evidence from a school reform in Bavaria," *Economics of Education Review*, 42, 12–33.

VILKAMA, K. AND V. BERNELIUS (2019): "Pikkulapsiperheiden muuttoliike muovaa koulujen oppilaspohjaa Helsingissä," *Kvartti - Kaupunkitiedon Verkkolehti*.