Incentives from curriculum tracking

Kristian Koerselman*

February 14, 2012

Abstract

Curriculum tracking creates incentives in the years before its start, and we should therefore expect test scores to be higher during those years. I estimate incentive effects using variation from policy experiments in the UK and Sweden as well as from an international cross-section. I find evidence for incentive effects of tracking in the UK and internationally, while Swedish results are inconclusive. Incentive effects of tracking show how early age scores can be endogenous with respect to later-age policies, and add to a growing literature on incentives in education.

Keywords: incentives, curriculum tracking, high-stakes testing, student achievement

*Helsinki Center of Economic Research, University of Helsinki. Contact information at http://economistatwork.com

1 Introduction

Curriculum tracking is the explicit separation of students into schools or classes based on observed past or expected future achievement. The tracking literature has mainly focused on the later-age effect of curriculum tracking on educational achievement and wages, measuring outcomes after the end of compulsory education or later. I argue that there are good reasons to look at the effects of tracking policies on early age student outcomes as well.

Tracking creates incentives before its start, amongst others for students to work harder in order to get into a higher track. The tracking point is a high-stakes moment for the student, whether the track choice is based on an explicit test or not.

The idea of incentive effects of tracking is not new. In some form or another it can for example be found in Galindo-Rueda and Vignoles (2004), Waldinger (2006) and Eisenkopf (2009). I add to this literature by making a comprehensive empirical analysis of the phenomenon using three different data sources. I find robust causal evidence for incentive effects in the UK and a strong correlation between tracking policies and early test scores in international data. Estimates based on a Swedish school reform are unfortunately inconclusive.

Incentive effects of tracking have two main implications. First, they illustrate that early age educational outcomes are endogenous with respect to later age educational policies. This means that we should not use test scores at a certain age to evaluate policies before that age without taking into account policies after that age, that we should not blindly use value-added specifications to measure the later age effects of policy, and that regressing pre-policy outcomes on policy does not generally make for a good 'placebo test' of post-treatment identification.

Second, there is a growing literature on incentives and high-stakes testing in education (e.g. Bishop 2006, Neal and Schanzenbach 2010, Juerges et al. 2012). We know that high-stakes at the end of middle or high school can lead to higher student test scores and sometimes achievement. The results presented in this paper add to this, and show that institutional incentives affect measured achievement at earlier ages as well.

2 Background

There are many mechanisms through which tracking can increase early test scores. The most direct incentive effect is through students. It pays for them to work harder before the tracking point in order to end up in the higher track. Attending the higher track will give students a better peer group, which will in turn increase future achievement. Upper track attendance will also usually leave open the possibility to enter higher education at the end of secondary school, and is a labor market signal of ability of its own. All these factors give the student an incentive to substitute effort towards the pre-tracking period.

Even if primary school students may not grasp the full consequences of their track placement, their parents will. To the degree that parents care about their children's outcomes, they will also have an increased incentive to aid their children's learning before the tracking point, and they are likely to push their children harder as well.

Teachers have an incentive to teach better. It seems a reasonable assumption that teachers should do this for their students' sake, but it may also be in their own interest to do so. The track placement of students (and the possible test preceding it) makes teacher quality more visible, and makes it easier for principals to reward and punish teacher effort as well as easier for parents to choose better schools for their children.

Students and teachers may substitute effort between subjects: from non-tested subjects to tested ones. Because of spillover effects between subjects, the net effect of tracking on achievement in non-tested subjects does however not have to be negative. (cf. Winters et al. 2008)

Tracking policies may also affect the early curricula and teaching styles in a more institutionalized way. The educational system may evolve towards stressing early achievement more, especially in tested subjects. Of course, the direction of causality can also run the other way if early achievement oriented regions have refrained from delaying the tracking point (cf. Betts 2010).

To at least some degree, incentive effects can cause students to do better at tests rather than learn more on an underlying level (cf. Klein et al. 2000, Jacob 2005, Almlund et al. 2011 section 5.6). This is a problem if we want to use incentives to increase underlying achievement. For the methodological implications of the endogeneity of test scores does however not matter whether endogenous test scores reflect endogenous learning or not.

Very good and very poor students may already feel certain of their future track placement regardless of the amount of effort they put in. One could therefore think that incentive effects should only occur just under the threshold for entering the higher track. Reality is however likely to be more complex. Direct peer effects may cause poor students to feel actively discouraged while students above the threshold may put in more effort when some of their classmates catch up. Indirect peer effects through parents, schools and teachers as well as through changes to the curriculum may affect all students, not only those below the threshold. Also, the selection process into the higher track is likely to be noisy and students are unlikely to be aware of their exact position relative to the threshold. For all these reasons we should expect to see incentive effects over the entire test score distribution, though not necessarily of the same sign or magnitude.

3 UK evidence for incentive effects

Since the Second World War, the UK has gradually gone from a tracked to a comprehensive school system. In the old system, students were split around age 11, after which they either entered an upper track grammar school, or a lower-track secondary modern, at least partly based on an achievement test. In the new system, all students attended a comprehensive school in order to make available to all children "all that is valuable in grammar school education" (Government Circular 10/65, 1965).

The Labour government had entered the 1964 elections with a promise to abolish the tracked educational system, and wanted to impose the new comprehensive system "as rapid as possible." Even so, the Labour government "requested" rather than demanded that Local Education Authorities (LEAs) change their policies, and the rate of change was initially limited.

	students	difference
full sample	18558	0
age 7 and 11 scores known	12066	-6492
age 11 LEA known	11098	-968
tracking status known	8114	-2984
tracking change not in 1969	7150	-964

Table 1: Number of observations in the full NCDS sample, as well as in subsamples with increasingly stringent inclusion conditions. The main reason for missing tracking information is students attending private schools.

The hesitant Labour attitude was induced by both practical and political concerns. On the one hand, extensive planning was needed in order to create the new schools, in part because of existing investment in school buildings. On the other hand, LEAs had had considerable autonomy in setting educational policies themselves since 1944, and their position was strengthened by the rather narrow Labour majority in parliament in combination with opposition against reform from within the Labour party.

In the end, comprehensive schools were implemented in a region-by-region, school-by-school fashion, both by merging or converting existing schools and by creating new ones. (Government Circular 10/65, 1965; Benn and Chitty 1996, ch. 1; Kerckhoff et al. 1996, ch. 2)

The survey most appropriate to study the UK reform is the longitudinal National Child Development Study (University of London 2008) or NCDS. It aims to follow all those born in Great Britain in the week starting on the 3rd of March 1958. The 1958 cohort turned 11 in 1969, when one part of the cohort was selected into one of two tracks, while the other part entered the comprehensive school system. I will use the 1958 sweep (at the time called Perinatal Mortality Survey) as well as the 1965, 1969 and 1974 sweeps, when the subjects were 0, 7, 11 and 16 years old.

As can be seen from Table 1, out of the full sample of 18558 students 11098 are left after we require age 7 and age 11 test scores as well as geographical information to be known. I treat the other 7460 as missing at random conditional on observables.

It is not a priori clear what tracking status should be assigned to private schools. I judge that to treat private schools as missing on the tracking variable is the more conservative choice, and will report estimates excluding this group below. A small number of private schools indicate that they are comprehensive in the survey. When I include these as comprehensive, and other private schools as tracked, the empirical results stay virtually unchanged.

Another 2984 students disappear from the sample when we exclude private schools and require tracking information to be known. I also disregard students whose schools turned comprehensive in the very year they took the age 11 test, because it is unclear what information they had on the status of their future schools. I have 7150 students left in the final sample.

The 1974 sweep of the NCDS recorded the tracking status and reform year of the school the individuals were attending at that point. This measure can be used to reconstruct the year of reform relative to 1969, the year the individuals entered the secondary school system.

The distribution of students exposed to the different reform years in the sample can be seen from Figure 1. The students on the left side of the figure entered a secondary school that had reformed before 1969, which means that the students entering them could be sure of its comprehensive status. Those on the right side entered a school that reformed only after 1969, i.e. after our cohort had entered them. Students may have had some information on the coming reform, but their subjective probability of entering a tracked system will have been smaller the later the reform actually took place. Students in the 'later' category were never part of a comprehensive school during their educational career.

There are multiple measures of age 11 achievement in the data: a general ability test containing both verbal and non-verbal items, a reading comprehension test and a mathematics/arithmetic test. In addition to these, we have teacher assessments of student abilities in different domains.

I synthesize all these variables into one in a two step process. First, I convert all test score distributions to z-scores because their shapes are arbitrary and skewed, and contain little cardinal level information on underlying achievement. Then, I extract the first principal component of the normalized scores to end up with a measure of general achievement. This process also has the advantage of reducing measurement error from any of the specific tests.

I calculate reliability ratios for both the age 7 and age 11 principal components under the assumption that all measurement error is white noise. This allows me to inflate the measures' standard deviations in such a way that the point estimates will be expressed in standard deviations of the signal. Because the reliability ratios are close to unity, the difference between this method and simply reporting effect sizes is small in practice.

I encode tracking status at age 11 (T_s) as a dummy indicating whether the student's school turned comprehensive before 1969, or after. I also select two groups of control variables, listed in Tables 4 and 5 starting on page 12. The first group A_i consists of standardized age 7 scores and teacher ratings. These include the results of a word recognition and word comprehension test, a copying designs test to assess perceptuo-motor abilities, a draw-a-man test to assess general mental and perceptual abilities, and an arithmetic test.

The second group X_i is a selection of a wide variety of parent and student background variables. I use six linearized measures of parental involvement in the child's education at age 7 and a large number of categorical parent and student background variables.

Unfortunately for our purposes, reforms were not implemented at random. Richer, right-wing areas were slower to reform (Benn and Chitty 1996, ch. 1, Galindo-Rueda and Vignoles 2004), and a simple comparison of tracked and comprehensive areas or schools is therefore likely to show incentive effects even if none exist in reality. Successful identification of the causal effect of tracking will have to come from adequately controlling for primary school inputs such as ability and parental and student background variables.

Additionally, there may be selection within and between regions due to non-compliance. Families with good students may move to a tracked area when faced with a comprehensive secondary school, while families with poor students may seek out comprehensive areas.

In areas where upper track schools remained, the new comprehensive school may in effect have become the new lower track school, with the upper track

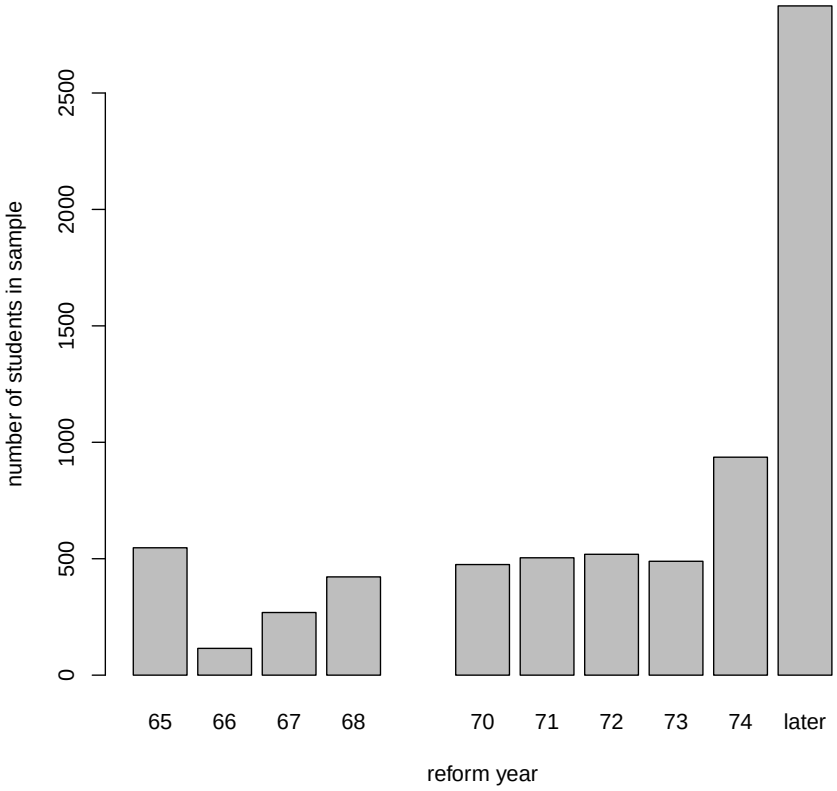

Figure 1: Number of students in sample by reform year. The students in the sample all turned 11 in 1969, at which point they were split into tracks in the pre-reform system. Those entering reformed secondary schools (reform year before 1969) should be expected to have lower age 11 scores than those entering schools that reformed after 1969.

Dependent variable: UK achievement age 11 (1969)						
specification	(1)	(2)	(3)	(4)	(5)	(6)
tracking (T)	0.15	0.10	0.10	0.10	0.10	0.08
	0.04	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*
ability (A_i)		yes	yes	yes	yes	yes
controls (X_i)			yes	yes	yes	yes
students	7150	7150	7150	5634	7150	7150
grouping	schools	schools	schools	schools	LEAs	years
groups	645	645	645	556	167	10

Table 2: Incentive effects in the UK. Students who knew their lower secondary school would be comprehensive score lower than those who had reason to expect a tracked school. Standard errors in italics.

school attracting all good pupils. Since we can control for ability and background, both forms of selection will lead to an overestimate of incentive effects only to the degree that movers are *unobservably* different in the expected direction. I will try to control for all these kinds of selection below.

To take into account the hierarchical nature of the data, I estimate a multilevel or hierarchical linear model (e.g. Gelman and Hill 2007, Pinheiro and Bates 2009) with regressors and error terms on different, nested levels. For the baseline regressions there are two levels: individuals, and LEA×reform year combinations, which I will henceforth call 'schools'.

In the first specification

$$y_{s,i} = \alpha + T_s\beta + \varepsilon_s + \varepsilon_i \tag{1}$$

individual achievement $y_{s,i}$ is regressed on a school level tracking variable T_s, and includes error terms both on the school and on the individual level.

Adding individual-level control matrices A_i and X_i allows us to explore the estimated effects of these background factors on an individual level, while retaining a school level estimate of the incentive effect of tracking.

$$y_{s,i} = \alpha + T_s\beta + A_i\gamma + \varepsilon_s + \varepsilon_i \tag{2}$$

$$y_{s,i} = \alpha + T_s\beta + A_i\gamma + X_i\delta + \varepsilon_s + \varepsilon_i \tag{3}$$

The estimates for these specifications can be seen from Table 2.

The first column shows that the unadjusted relationship between the tracking variable and age 11 scores is 0.15 of a UK standard deviation. This is a sizable difference, but probably an overestimate of the causal effect.

Turning to column (2), we can see that the estimated effect indeed declines to 0.10 when we control for age 7 scores. If we are lucky, the inclusion of age 7 test scores is enough to control for the nonrandom nature of the tracking reforms. In column (3), I have added all background variables in X_i as well. The estimate remains at 0.10. This strongly suggests that age 7 test scores pick up most of the selection.

Even if we can control for the non-randomness of reform areas, we are still left with possible problems of student selection between and within areas. I rerun specification (3) to include only students that did not move to a different LEA between ages 7 and 11. This reduces the number of students from 7150 to

5634, and the number of schools from 645 to 556 (the sampling method causes individual schools to be represented by small numbers of students). As can be seen from column (4), the estimate still stands at 0.10.

Next, I look at possible selection within areas by using the share of students exposed to a tracked school within each area as the measure of tracking for each student. I define an area as the Local Education Authority: the policy-setting authority of which there are 167 in the sample. As can be seen from column (5), the point estimate is still unchanged. Both results suggest that noncompliance is not a problem given the controls available to us.

As an additional check, I group all schools together by reform year, and define tracking as a year-level indicator variable.

$$y_{y,i} = \alpha + T_y\beta + A_i\gamma + X_i\delta + \varepsilon_y + \varepsilon_i \tag{6}$$

Even with a low number of year observations, the tracking estimate is still significantly different from zero, at a slightly lower point estimate of 0.08 because the results are now weighted by year rather than by school.

An illustration of this specification can be seen from Figure 2. The students on the left side of the figure knew they were going to enter a comprehensive school while those on the right side did not. We can speculate that those attending schools that reformed in later years had both less uncertainty over the continued tracked status of their secondary school and a larger actual incentive to enter the higher track. Such a pattern is indeed visible in the figure. Early test scores are not only larger on average for those entering a tracked school, but are also increasing in the number of years the reform lies in the future for any particular student.

It is important to remember that there is no a priori reason to expect a sharp discontinuity between students entering a school that had just reformed (i.e. in 1968), or was just about to reform (in 1970). Students entering schools that reformed in 1970 probably experienced much smaller incentives than those managing to exit lower secondary before the reform took place in their school.

I also estimate a model with age 7 achievement as the dependent variable as a kind of placebo test. Since we cannot control for early age scores when using them as the dependent variable, we should expect these estimates to include some selection. Still, as can be seen from Table 3, the estimated treatment effect is much smaller and not significantly different from zero for age 7 outcomes. Because I have inflated test scores to account for measurement error, this result is unlikely to be due to the age 7 measurements being more noisy. I interpret the results of the test as additional evidence for the credibility of the original specification.

Do incentive effects differ by gender or background? I add an interaction with gender to specification (3). Estimated incentive effects are larger for boys, but not significantly so. I also add interactions on father's socioeconomic status, but the uncertainty of the interactions is large. They are however suggestive of larger incentive effects among high-SES students. I have illustrated these results in Figure 3.

A quantile regression version of specification (3) suggests that incentive effects are slightly larger at the higher end of the distribution. Such a result would be intuitive if those at the lower end felt they had little hope of getting into the upper track in any case. This difference is however not statistically significant either. To illustrate this, I have used quantile regression on 100 bootstrap

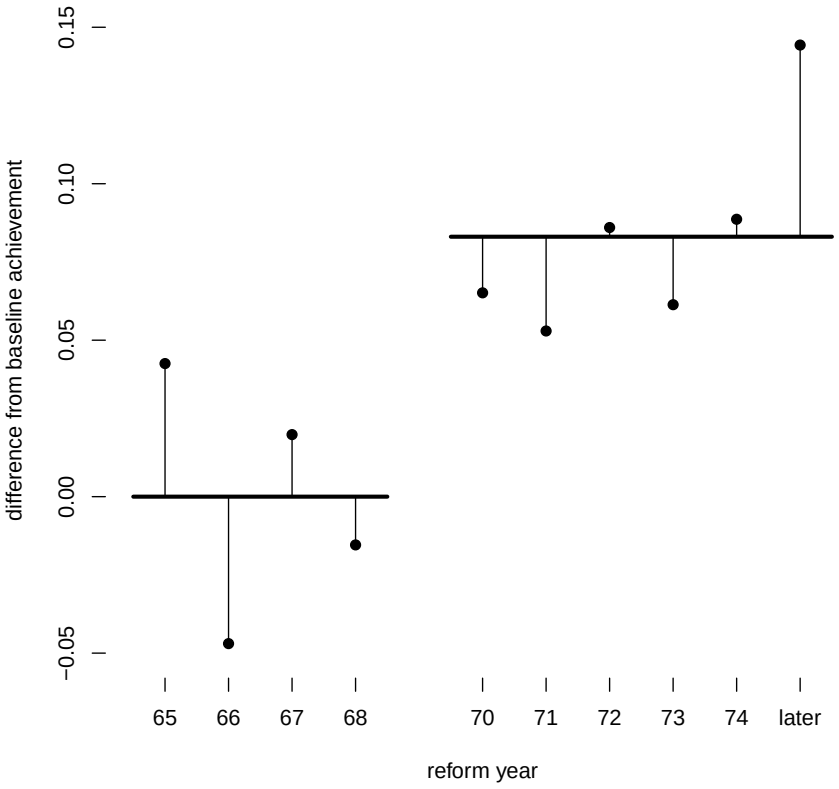

Figure 2: Secondary schools left of the divide turned comprehensive before the NCDS students could enter them. Achievement estimates from specification (6). Dots indicate the year-level errors.

dependent variable	age 11	age 7
specification	(7)	(8)
tracking (T)	0.13	0.04
	0.03	*0.03*
controls (X_i)	yes	yes
students	7150	7150
grouping	schools	schools
groups	645	645

Table 3: Placebo test for UK incentive effects using early age scores. Standard errors in italics.

replications of the data, subtracted the estimated effect on the median in every replication, and plotted the 5th, 50th and 95th percentile estimate for every test score quantile in Figure 4. This produces an indication of confidence bounds on the slope rather than on the location of the quantile profile.

It would be interesting to know more about the mechanisms through which incentives work on students. I try to see whether incentives seem to affect parental effort at age 11. The point estimate is positive, but again the power of the estimate is not high enough to reject a zero effect.

Summarizing, incentive effects look credible in the UK setting. The biggest threats to identification are the non-random nature of changes in tracking policies as well as noncompliance by parents and students. The estimated effect of tracking on achievement growth between ages 7 and 11 is however virtually unchanged when we add background variables as controls, lending credibility to the identification strategy. Neither excluding movers nor using LEA-level tracking variables change the point estimate much. Conclusions are even robust to grouping observations per reform year rather than by school, and survive an early-age placebo test.

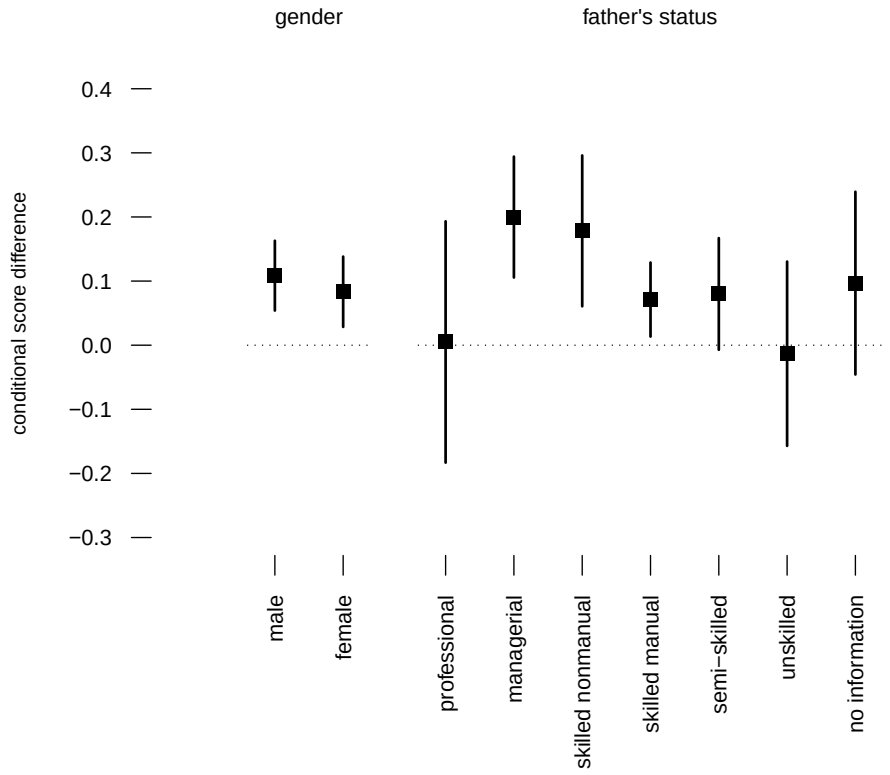

Figure 3: Estimated incentive effects for different subgroups. Bars indicate the 95% confidence interval. The size of the effect is not significantly different between boys and girls. Results are indicative of larger incentive effects for high SES children.

variable name	overall mean	overall sd	tracked mean	compr. mean
dependent variable y				
Achievement age 11	0.06	1.02	0.10	-0.13
Achievement age 7	0.05	1.03	0.08	-0.06
early ability A_i				
Arithmetic score age 7	0.00	1.00	0.01	-0.06
Copying designs score age 7	0.00	1.00	0.01	-0.05
Drawing score age 7	0.00	1.00	0.01	-0.02
Reading score age 7	0.00	1.00	0.03	-0.13
Creativity rating age 7	0.00	1.00	0.01	-0.06
Numbers rating age 7	0.00	1.00	0.03	-0.11
Oral ability rating age 7	0.00	1.00	0.01	-0.05
Reading rating age 7	0.00	1.00	0.03	-0.12
World awareness rating age 7	0.00	1.00	0.02	-0.09
early parental effort, in X_i				

continued on next page

continued from previous page

variable name	overall mean	overall sd	tracked mean	compr. mean
Father reads to child	1.07	0.77	1.08	1.04
Mother reads to child	1.31	0.72	1.32	1.27
Parents' initiative to discuss child with teacher	0.57	0.50	0.57	0.55
Father's interest in child's education	1.09	0.62	1.10	1.04
Mother's interest in child's education	1.22	0.68	1.23	1.16
Parents help school	0.54	0.50	0.55	0.49

Table 4: NCDS descriptive statistics: test scores and parental involvement in the child's education at age 7.

Table 5: NCDS student-weighted descriptive statistics: parent and background controls.

variable name	overall mean	tracked mean	compr. mean
Parent and student background controls, in X_i			
Gender			
male	0.51	0.51	0.50
female	0.49	0.49	0.50
Height quantile group age 11			
5	0.19	0.19	0.18
4	0.19	0.19	0.18
3	0.19	0.19	0.18
2	0.19	0.19	0.17
1	0.19	0.18	0.21
no information	0.07	0.07	0.08
Father figure			
natural	0.91	0.91	0.90
other or no information	0.09	0.09	0.10
Socio-economic status father			
professional	0.04	0.04	0.03
manegerial technical	0.16	0.17	0.14
skilled nonmanual	0.09	0.09	0.09
skilled manual	0.42	0.42	0.44
semi-skilled	0.16	0.16	0.17
unskilled	0.05	0.05	0.06
no information	0.06	0.06	0.06
Education father (ISCED)			
5	0.03	0.03	0.02
3	0.16	0.17	0.15
2	0.54	0.55	0.52
1	0.01	0.01	0.02
no information	0.25	0.25	0.29
Education mother (ISCED)			
5	0.02	0.02	0.01
3	0.19	0.19	0.19
2	0.57	0.57	0.56
1	0.01	0.01	0.01
no information	0.21	0.20	0.23
Father born			
British Isles	0.90	0.90	0.88
Eire or Ulster	0.03	0.03	0.03
other or unknown	0.07	0.07	0.08
Mother born			
British Isles	0.91	0.92	0.89
Eire or Ulster	0.03	0.03	0.03
other or unknown	0.06	0.05	0.08
Father reads books			
often	0.46	0.47	0.42
occasionally	0.20	0.19	0.23
hardly ever	0.27	0.27	0.26
no information	0.07	0.07	0.08
Mother reads books			
often	0.32	0.32	0.29
occasionally	0.21	0.21	0.22
hardly ever	0.42	0.41	0.44

continued on next page

continued from previous page

variable name	overall mean	tracked mean	compr. mean
no information	0.05	0.05	0.05
Accommodation type			
house	0.86	0.86	0.84
flat	0.07	0.07	0.08
rooms	0.02	0.01	0.02
other or no information	0.05	0.05	0.06
Child goes reluctantly to school age 7			
no	0.86	0.86	0.86
yes	0.10	0.10	0.10
no information	0.04	0.04	0.04
Poor at English age 7			
no	0.97	0.98	0.95
somewhat	0.01	0.01	0.02
certainly	0.00	0.00	0.01
no information	0.01	0.01	0.02

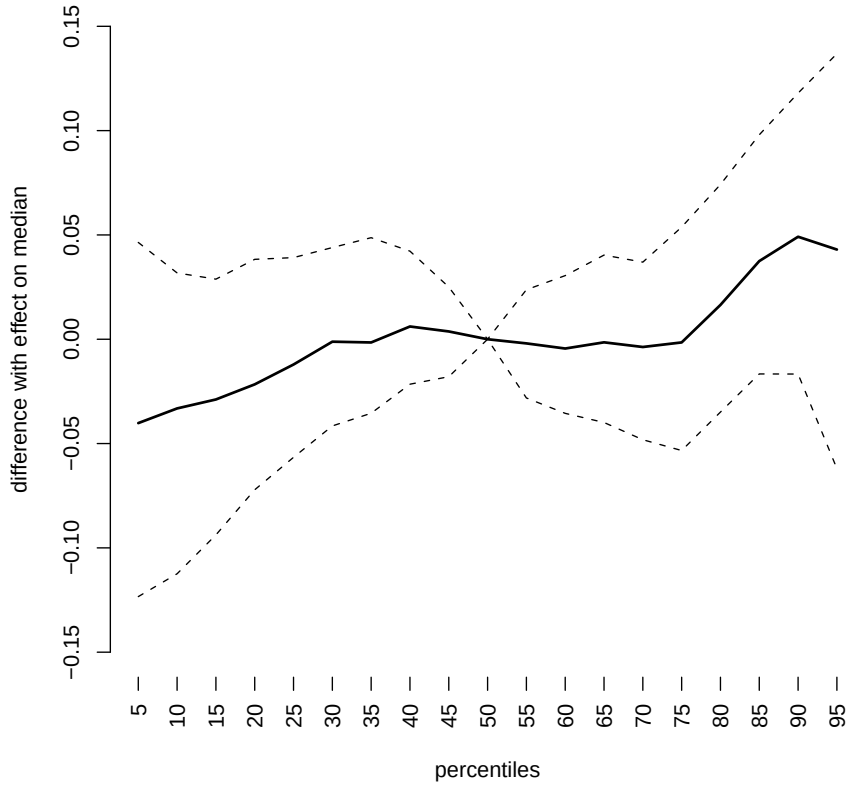

Figure 4: Estimated incentive effects relative to the effect at the median for specification (3). The dashed lines indicate approximate an 95% confidence interval for the slope of the quantile profile. The null hypothesis that the size of the effect is equally large as the effect on the median cannot be rejected.

4 Incentives in the Swedish comprehensive school reform

In the Sweden of the 1940s, there was a widespread feeling that that the educational system was inadequate for the country's needs. It was increasingly difficult to enter one of the limited number of upper track lower secondary schools, and this problem was only to increase when the big cohorts born immediately after the war were to enter secondary education.

The lower track was felt to be lacking as well. Other countries had been increasing the length of compulsory education, and Sweden was seen as falling behind. At the same time, the educational system was becoming a tool for the emancipation both of women and of the rural areas. It was also to foster democratic values, not by indoctrination but by "promoting respect for truth and the motivation to find it." (Statens Offentliga Utredningar 1948, p. 3)

While there was general agreement that the educational system needed to be improved, the question of whether tracking should be postponed at the same time led to intense debate. In 1950, parliament reached an agreement first to implement a comprehensive school in a select number of municipalities only. These schools were experimental, and had varying degrees of within-school tracking (Marklund 1981).

In 1962 parliament accepted the general implementation of the nine-year comprehensive secondary school, with within-school differentiation only in the 9th grade, even if within-subject differentiation continued to exist at earlier ages. (Marklund 1980, 1982, Richardson 1977/2004)

Sweden moved from a patchwork of schools and systems, many of them underresourced, to a single compulsory, comprehensive school. This changed the curriculum, the quality of education and its quantity. In the new system, families also received additional financial support now that they had to keep their children longer in school. (Marklund 1981)

It is important to stress that the reform also involved changes in the first six grades of primary school. The amount of English teaching was increased in part at the cost of Swedish. Though perhaps concentrated mainly in the years immediately following the 1950 decision, there was also experimentation with new teaching methods, involving less frontal instruction. (Marklund 1981)

It is not a priori self-evident how early incentives were affected by the Swedish comprehensive school reform. On the one hand, students competing for the upper track in the old system lost an early incentive when early selection was replaced with a later and softer selection mechanism. On the other hand, later educational opportunities may have increased for many, increasing the option value of continued effort.

I use the first two cohorts of the longitudinal Evaluation Through Follow-up studies (Swedish abbreviation: UGU) collected by the Department of Pedagogics at the University of Gothenburg and Statistics Sweden (see Harnqvist 2000) to see whether early test scores changed as a result of the reform. The surveys aimed to interview all born in Sweden on the 5th, 15th and 25th of each month in 1948 and 1953. The proportion of students for which background information is available is very high. For the 1948 cohort, the proportion of the target population for which background information is known is 98%. For the 1953 cohort this number is somewhat lower at 93% due to limited resources at Statistics

	students	municipalities
full sample	21877	1020
with IQ scores	19946	1013
with IQ and math scores	17427	1005
..of which tracked in 1948	8277	801
..of which comprehensive in 1948	1013	145
..of which tracked in 1953	1643	313
..of which comprehensive in 1953	6494	617

Table 6: Number of observations in the full UGU 1948 and 1953 sample, as well as in the subsample with known ability scores and ability and mathematics scores respectively. As can be seen from the last four rows, the panel of municipalities is not balanced.

Sweden at the time.

The majority of the 1948 cohort was in 6th grade in the academic year starting in 1960, at a time when experimentation with comprehensive schools was fully underway. When the 1953 cohort entered 6th grade in 1965, the comprehensive school had been implemented in most, but not all municipalities.

I have data on spatial, verbal and inductive components of an age 12 ability test for most students, as well as standardized tests in mathematics for those who were in 6th grade of primary school. Like before, I transform each ability subscale into a standard normal distribution, take their first principal component and inflate it so that the standard deviation of the latent trait is one. I transform the math score distribution into a standard normal distribution as well, but unfortunately, I do not have enough information on subscores to estimate reliability ratios.

I have at least some information for 21877 students in 1020 municipalities in the full sample. As can be seen from Table 6, this decreases to 19946 students in 1013 municipalities for which I have information on IQ, and further to 17427 students in 1005 municipalities for those which I have math scores as well.

While it may not be all too far from the truth that the students without IQ scores were missing at random conditional on covariates, the students with IQ scores but without a mathematics test score are not a random selection. They partly consist of those that either were not in 6th grade when their peers were, and of those that had transferred to an upper track school at an earlier age. I will look at the effects of excluding this group further below.

Municipality×cohort level means and standard deviations of all included variables can be found in Table 7.

I define a municipality as tracking if at least one student in the municipality is reported to be in a tracked school. According to this definition, 85% of municpalities in the final sample were tracked in 1960 and 34% were in 1965.

I estimate variations of a fixed effects model

$$y_i = \alpha + T_i\beta + MC_i\gamma + X_i\delta + Z_i\zeta + \varepsilon_i \tag{10}$$

where y_i is an ability or achievement outcome, T_i is municipal tracking status, MC_i is a matrix of municipality and cohort indicators, X_i is a matrix with municipality×cohort background variables, Z_i is a matrix with individual background variables, and ε_i is the error term. I weight individual observations

	mean	sd
IQ	0.03	0.53
math	-0.06	0.56
tracking	0.59	0.49
female	0.50	0.27
month of birth	6.34	1.80
father's education primary	0.87	0.19
father's education lower secondary	0.06	0.13
father's education upper secondary	0.04	0.10
father's education university	0.02	0.07
father's education unknown	0.02	0.06
mother's education primary	0.88	0.19
mother's education lower secondary	0.08	0.15
mother's education upper secondary	0.02	0.08
mother's education university	0.00	0.02
mother's education unknown	0.01	0.07
municipality×cohort sample size	9.29	31.02

Table 7: Sample means and standard deviations by municipality×cohort for the subsample for which both IQ and math scores are known.

Dependent variable:	IQ (9)	math (10)	math (11)	IQ (12)
early tracking	-0.07	-0.05	0.00	-0.02
	0.03	*0.04*	*0.04*	*0.03*
ability controls			yes	
other controls	yes	yes	yes	yes
students	17427	17427	17427	19946
groups	1864	1864	1864	1919

Table 8: Estimates of the effects of the Swedish comprehensive school reform on early test scores. Standard errors in italics.

with the inverse of the number of observations per municipality×cohort, and use standard errors clustered on the municipality×cohort level.

I have listed estimation results in Table 8. As can be seen from column (9), there seems to be a significantly negative conditional relationship between tracking and IQ, while we can see from column (10) that the coefficient on math scores is negative, but not significantly different from zero.

The apparent effect on IQ seems implausibly large. For example, Pekkarinen et al. (2009) find effects on later age military test scores an order of magnitude smaller than these. Even if we believe that incentive effects are stronger than the later age effects of tracking, how can it be possible that policy has a larger effect on IQ than on mathematics?

One explanation could be that the math scores have large amounts of measurement error. Unfortunately, there is not enough information in the UGU data set to check for this.

Another possibility is that the sample is not representative of the student population in each municipality. Mathematics scores are only known for those students who were in the 6th grade of either the new comprehensive school or of the old primary school, leading to biased estimates.

I rerun the IQ regression of the second column on a sample including the students with missing mathematics scores. As can be seen from column (12),

selectively missing students within municipalities seem to be able to explain most of the negative conditional correlation between tracking and IQ.

Selectively missing individuals are an argument in favor of controlling for IQ as it identifies a mechanism by which IQ scores can be conditionally correlated with tracking even if there is no causal effect of the reform on IQ.

Under the assumption that the true effect of tracking on IQ is zero, we can use IQ to control for selection. As can be seen from column (11), the estimated effect of the reform on math scores conditional on ability scores is very close to zero.

The best estimate of the reform effect on IQ comes from column (12), and the best estimate of the effect on mathematics skills comes either from column (10) or (11), depending on assumptions. None of these three estimates is significantly different from zero. It is possible to obtain borderline significantly positive or negative effects with other model variations, but these results are never robust to small and arbitrary changes.

One could speculate that the lack of clear results are due to measurement error in the reform variable. To check on this, I merge the UGU data with reform year data which Holmlund (2007) has collected. I obtain point estimates close to the estimates in Table 8, and I conclude that measurement error in the reform measure is not likely to be the main driver of these results.

In a bid to increase efficiency, I consider an alternative family of models which uses county and cohort fixed effects, with separate intercepts for the three largest cities, and municipality×cohort random effects. These random effects models can be more efficient, but cannot control for potential municipality level selection. I test for bias in the random effects models using a Hausman specification test under the assumption that the fixed effects model is consistent. For most specifications I reject the null that the random effects model is consistent at the 5% level, and I conclude that it cannot be used to improve the previous results.

The Swedish comprehensive school reform changed many aspects of education simultaneously, and what we measure are the combined effects of multiple mechanisms. The reform involved many changes: to the pre-test curriculum and perhaps also to pre-test teaching styles, to the option value of continued education and to its cost, and also to the amount of compulsory education. It is possible that what we are measuring is a positive incentive effect of tracking canceled out by a combination of changing general incentives and improved early age learning. In this respect, the British reform is a much cleaner policy experiment than the Swedish one.

5 International evidence for incentive effects

The International Association for the Evaluation of Educational Achievement administers various standardized tests in a large number of countries. This allows us to look for incentive effects cross-sectionally. I use two waves of two of the most well-known studies: the Trends in International Mathematics and Science Study TIMSS, and the Progress in International Reading Literacy Study (IEA 1995, 2001, 2003, 2006). PIRLS is an internationally comparable early age reading literacy survey. TIMSS surveys mathematics and science literacy at three different grades, of which I use the earliest. Both surveys aim to test

a representative sample of the population of fourth graders in the participating countries. I take the average of TIMSS mathematics and science scores to get a more general measure of achievement.

I make no attempts to estimate measurement error in these data, and I standardize the achievement measures to have standard deviation one in the student population in my sample. Rindermann (2007) finds high correlations between country means in international achievement surveys. This is an indication that measurement problems in international surveys are perhaps not as large as one could otherwise think, at least when it comes to country means.

I take tracking information mainly from the Eurybase database (Eurydice 2008), supplemented with information from Wikipedia and from various countries' ministry of education websites. I drop a small number of nonwestern countries with conflicting information on tracking policies. The tracking variable I will use is the age at which a substantial proportion of students will be tracked into different schools. This definition is close to that of Hanushek and Woessmann (2006). Even though I try to pinpoint the start of tracking in each country to an exact age, I use a dummy variable in the analysis, indicating tracking at an age of 12 or earlier. Though this seems somewhat arbitrary, it is not more so than to assume that incentive effects would be linear in years. Nevertheless, results are robust to using a different cutoff, or using a continuous tracking age instead.

As control variables, I use real per capita purchasing power-adjusted GDP (expressed in 10 000 USD) from the Penn World Table (2006) as well as educational expenditures as a percentage of GDP from the World Bank EdStat database (2011). For GDP, the year of observation is always 1995. For educational expenditures, it is the available observation the closest to 1995. Descriptive statistics for these and other variables can be seen from Table 9. I have complete data on 1040596 students in 51 countries.

A more useful sample is probably the subset of countries in the original sample that is a member of the European Economic Area or EEA. Not only is the EEA a more homogeneous group of countries, reducing omitted variable bias, my tracking measure used is most relevant in a European context, as it classifies within-school tracking countries as late tracking (Betts 2010). This reduces the sample to 515788 students in 28 countries.

As in Section 3, I estimate a multilevel model to take into account the errors individuals have in common when they share a class, school or country. The error structure in all specifications is nested, and given by

$$\varepsilon \equiv \varepsilon_{cn} + \varepsilon_s + \varepsilon_{cl} + \varepsilon_i$$

where subscripts cn, s, cl and i stand for country, school, class and individual respectively.

The first specification gives the raw relationship between individual scores $y_{cn,s,cl,i}$, and the country-level tracking regime T_{cn}. The multilevel model takes care of the difference in levels in its calculation of standard errors of the various parameter estimates. I add an variable D_s, indicating whether the score is a PIRLS or a TIMSS score.

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + \varepsilon \tag{11}$$

The results from estimating this equation can be seen from column (11) in Table 10. Countries with early tracking clearly have higher score means, with

variable	weighting by student μ	by student σ	by country μ	by country σ
Full sample:				
test score	0.00	1.00	0.13	0.89
per capita GDP ('0 000 1995 USD)	1.46	0.99	1.41	0.82
educational expenditures (%GDP)	4.52	1.32	4.99	1.58
books at home	0.31		0.32	
female	0.47		0.48	
students				1040596
countries				51
European Economic Area only:				
test score	0.41	0.68	0.30	0.68
per capita GDP ('0 000 1995 USD)	1.75	0.53	1.54	0.68
educational expenditures (%GDP)	4.96	0.91	5.23	1.33
books at home	0.34		0.35	
female	0.50		0.49	
students				515788
countries				28

Table 9: International data: descriptive statistics for the full sample (top), and for the EEA countries only (bottom).

Dependent variable: international early age achievement					
	(11)	(12)	(13)	(14)	(15)
tracking (T)	0.41	0.17	0.23	0.23	0.26
	0.19	*0.16*	*0.07*	*0.07*	*0.07*
GDP		0.38	-0.01	-0.02	-0.01
		0.08	*0.05*	*0.05*	*0.05*
expenditures		-0.08	0.03	0.02	0.03
		0.04	*0.03*	*0.02*	*0.03*
books at home				0.14	
				0.00	
$T\times$books at home				0.00	
				0.04	
female					0.04
					0.00
$T\times$female					-0.05
					0.02
students	1040596	1040596	515788	515788	515788
countries	51	51	28	28	28

Table 10: International evidence for incentive effects; pooled multilevel regression based on international data. Standard errors in italics.

the mean difference as large as 0.41 standard deviations of international student test scores.

There is no reason to assume that the estimated effect is not due to some third factor. This becomes apparent when we add real per capita GDP and educational expenditure as controls in the next specification. Both variables are contained in the country level matrix C_{cn}.

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + C_{cn}\delta + \varepsilon \tag{12}$$

The estimates from this specification can be seen from column 12. Estimated incentive effects are now more than halved at 0.17 standard deviations.

However when we turn to the EEA sample, the estimate is improved in many ways. It can be seen from column (13). GDP and educational expenditures now play a much smaller role, both turning statistically insignificant.

At 0.23, the estimated incentive effects are now larger, but also much more precisely estimated. This is exactly what we should expect if the tracking variable has classical measurement error for non-EEA countries.

I have illustrated the estimate from specification (13) in Figure 5. As can be seen from the figure, a specification linear in age may seem to fit the data better, but the results would become more sensitive to the exact tracking ages we assign to late tracking countries.

The estimate is still not likely to reflect a causal effect in the sense that a country that randomly decides to change its tracking policies is likely not to experience a change in early test scores as large as 0.23 international standard deviations. One can easily imagine that the pattern is a combination of incentive effects of tracking, and a tendency for countries that stress the importance of achievement on hard, testable subjects in primary school to have retained a tracked secondary school system. The remarkably strong pattern in Figure 5 does however suggest that early tracking and early achievement are strongly related.

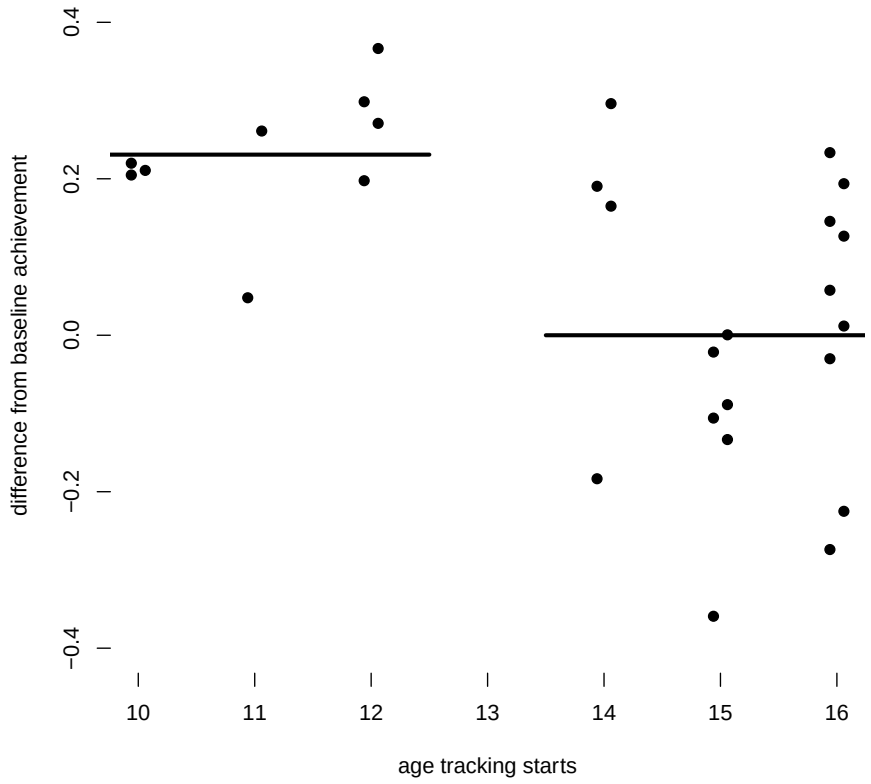

Figure 5: An illustration of the EEA estimate of incentive effects from specification (13). Early tracking countries have higher conditional early test scores. The solid line represents the estimate, dots indicate the country-level errors. The horizontal axis has been jittered slightly to improve visibility.

I estimate whether estimated effects differ for children with different parental backgrounds. For this, I use a dummy variable B_i which indicates whether the student has one case of books or more at home, the only SES variable that is available for all four surveys.

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + C_{cn}\delta + B_i\theta + (B_i \cdot T_{cn})\kappa + \varepsilon \tag{14}$$

Because this specification includes an interaction between variables on two different levels, I bootstrap the standard error for the interaction term.

Results can be seen from column (14). Students with more than one case of books at home score higher on average, but the interaction with tracking is insignificant and close to zero.

In the last specification, I check whether the effects are different for boys than for girls. F_i is a dummy variable indicating whether the individual is female.

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + C_{cn}\delta + F_i\lambda + (F_i \cdot T_{cn})\mu + \varepsilon \tag{15}$$

Looking at column (15) of Table 10, we can see that the differences between boys and girls are moderately small at -0.05. Both the unclear differences in parental background and the smaller point estimate of incentive effects for girls mirror the UK findings.

Hanushek and Woessmann make a slightly different assessment of the tracking age, even if they define tracking in the same way. A re-run of my regressions with an age 14 tracking dummy based on the Hanushek and Woessmann variable gives higher and more precise point estimates in specifications (11) and (12), but makes no difference in the EEA sample of the later specifications.

All in all, international test score data provide us with an additional line of evidence for incentive effects. The estimated effect is unlikely to reflect an unidirectional causal link between tracking and early test scores only, but the relationship is nevertheless exceptionally clear.

6 Discussion

Given economic intuition as well as previous empirical research on high-stakes testing, it should be expected that tracking has an incentive effect on test scores before its start; parents, teachers and students should all be expected to respond to the incentives created.

In this paper, I find empirical evidence to support this hypothesis. In UK data, tracking causes an incentive effect of 0.10 UK standard deviations. Within the European Economic Area, tracking is associated with 0.23 international standard deviations higher scores. These estimates are large, but no larger than the 0.2–0.3 Jacob (2005) finds for a high-stakes test.

While it is hard to interpret the results of the international analysis causally on their own, they show a remarkable pattern, and add a line of evidence to the UK results, where the effect seems robust and well-identified. Swedish results are unfortunately inconclusive, probably both because of selectively missing data and because the Swedish comprehensive school reform consisted of many simultaneous policy changes.

Incentive effects of tracking have a number of implications. First, they illustrate that early age educational outcomes are endogenous with respect to

later age educational policies. Individuals are forward-looking, and measured outcomes are a result of policies at both earlier and later ages than the age of measurement. Consequently, we should not use test scores at a certain age to evaluate policies before that age without taking into account policies after that age as well.

Methodological implications extend to analyses where early test scores are not themselves the outcome of interest. Value added specifications are regularly used to control for unobservables (see e.g. Todd and Wolpin 2003). Such specifications can lead to biased estimates if the early age outcomes are affected by the policy under consideration.

For example, Hanushek and Woessmann (2006) use pre-tracking achievement to control for unobservables in their analysis of later age achievement, and find a negative effect of tracking on mean test scores. If we believe in incentive effects of tracking, their specification is invalid, and leads to downward biased estimates. The authors find an effect not significantly different from zero when omitting early scores from their specification.

The existence of incentive effects can invalidate the use of early outcomes in some 'placebo tests' as well. In a carefully controlled experiment, we may expect to find no difference between pre-treatment outcomes in treatment and control groups. In the case of natural experiments, subjects may be aware of their future treatment status, and act on it. For example, Manning and Pischke (2006) reject UK studies on tracking because they find that test score growth between age 7 and 11 is correlated with tracking policies after the age of 11. I argue that this correlation is exactly what we should expect.

Second, incentive effects on tracking also add a line of evidence to the literature on incentives in education. There are clear parallels between the start of tracking in early tracking systems on the one hand, and the minimum competency exams and curriculum-based external exit exams that are commonly held a few years later (e.g. Bishop 2006, section 3; Juerges et al. 2012). The results presented in this paper show that this kind of incentives are not only important at the end middle or high school, but can also affect outcomes at the end of primary school.

Even so, it is not clear that early tracking is a good instrument to increase competition and incentives in schools. Early tracking has a cost associated with it in terms of inequality and probably also in intergenerational mobility, and its later age positive effects on learning may not be very large. Increased incentives of tracking can also have more direct negative effects on intrinsic motivation and well-being (cf. Juerges et al. 2012), and we may actually want to delay tracking in places where primary school children are already under high pressure to achieve.

Acknowledgments

I thank Tuomas Pekkarinen, Markus Jäntti, Ludger Woessmann, Heikki Kauppi, Jonas Lagerström, Sari Kerr, Roope Uusitalo, Elias Einiö, Sangjun Jeong, Fabian Pfeffer, Sharon Simonton, Anders Stenberg, John Micklewright, Oskar Nordström Skans as well as participants at various conferences and workshops for their help and advice. I gratefully acknowledge financial support from *Yrjö Jahnssonin säätiö, Stiftelsens för Åbo Akademi forskningsinstitut*,

Bröderna Lars och Ernst Krogius forskningsfond, *Åbo Akademis jubileumsfond*, and from the *Academy of Finland*.

References

M. Almlund, A. Duckworth, J. Heckman, and T. Kautz. Personality psychology and economics. volume 4 of *Handbook of the Economics of Education*, pages 1 – 181. Elsevier, 2011.

C. Benn and C. Chitty. *Thirty years on: is comprehensive education alive and well or struggling to survive?* David Fulton Publishers, 1996.

J. Betts. The economics of tracking in education. *Handbook of the Economics of Education*, 3, 2010.

J. Bishop. Drinking from the fountain of knowledge: Student incentive to study and learn-externalities, information problems and peer pressure. *Handbook of the Economics of Education*, 2:909–944, 2006.

G. Eisenkopf. Student Selection and Incentives. *Zeitschrift fur Betriebswirtschaft*, 79(5):563–577, 2009.

Eurydice information network on education in Europe. Eurybase database on education systems in Europe. http://www.eurydice.org, 2008.

F. Galindo-Rueda and A. Vignoles. The heterogeneous effect of selection in secondary schools: understanding the changing role of ability. IZA discussion paper no. 1245, 2004.

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

E. Hanushek and L. Woessmann. Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, 116:C63–C76, 2006.

H. Holmlund. A researcher's guide to the Swedish compulsory school reform. Swedish Institute for Social Research (SOFI) Working Paper 9/2007, 2007.

K. Härnqvist. Evaluation through follow-up. A longitudinal program for studying education and career development. I C.-G. Janson (Red.). *Seven Swedish longitudinal studies in behavioural science*, 2000. Distributer: Swedish National Data Service (SND).

IEA. Trends in International Mathematics and Science Study TIMSS. 1995.

IEA. Progress in International Reading Literacy Study PIRLS. 2001.

IEA. Trends in International Mathematics and Science Study TIMSS. 2003.

IEA. Progress in International Reading Literacy Study PIRLS. 2006.

B. Jacob. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6): 761–796, 2005.

H. Juerges, K. Schneider, M. Senkbeil, and C.H. Carstensen. Assessment drives learning. the effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31(1), 2012.

A. Kerckhoff, K. Fogelman, D. Crook, and D. Reeder. *Going comprehensive in England and Wales: a study of uneven change*. Woburn Press, 1996.

S. Klein, L. Hamilton, D. McCaffrey, and B. Stecher. What do test scores in Texas tell us. *Education Policy Analysis Archives*, 8(49):1–22, 2000.

A. Manning and J. Pischke. Comprehensive versus selective schooling in England and Wales: what do we know? NBER working paper no. 12176, 2006.

S. Marklund. *Skolsverige 1950-1975. Del 1. 1950 års reformbeslut*. Liber UtbildningsFörlaget, 1980.

S. Marklund. *Skolsverige 1950-1975. Del 2. Försöksverksamheten*. Liber UtbildningsFörlaget, 1981.

S. Marklund. *Skolsverige 1950-1975. Del 3. Från Visbykompromissen till SIA*. Liber UtbildningsFörlaget, 1982.

D. Neal and D. Schanzenbach. Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2): 263–283, 2010.

T. Pekkarinen, R. Uusitalo, and S. Kerr. School tracking and development of cognitive skills. VATT working paper 2, 2009.

J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer Verlag, 2009.

PWT. Penn world table version 6.2. Alan Heston, Robert Summers and Bettina Aten; Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, September 2006.

G. Richardson. *Svensk utbildningshistoria: skola och samhälle förr och nu*. Studentlitteratur, 1977/2004.

H. Rindermann. The g-factor of international cognitive ability comparisons: The homogeneity of results in pisa, timss, pirls and iq-tests across nations. *European Journal of Personality*, 21(5):667–706, 2007.

Statens Offentliga Utredningar. 1946 års skolkommissions betänkande med förslag till riktlinjer för det svenska skolväsendets utveckling. 1948:27, 1948.

P. Todd and K. Wolpin. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113:F3–F33, 2003.

UK Department of Education and Science. Circular 10/65. United Kingdom, 1965.

University of London. Institute of Education. Centre for Longitudinal Studies. National Child Development Study: Local Authority Data, 1958-1974: Special Licence Access [computer file]. 2nd Edition. Colchester, Essex: UK Data Archive [distributor], August 2008. SN: 5744, 2008.

F. Waldinger. Does tracking affect the importance of family background on students' test score. Unpublished manuscript, LSE, January 2006.

M. Winters, J. Greene, and J. Trivitt. The impact of high-stakes testing on student proficiency in low-stakes subjects. Manhattan institute for policy research, Civic report no. 54, 2008.

World Bank. EdStat Education Statistics. 2011.