Swedish Institute for Social Research (SOFI)

Stockholm University

WORKING PAPER 3/2011

INCENTIVES FROM CURRICULUM TRACKING: CROSS-NATIONAL AND UK EVIDENCE

by

Kristian Koerselman

Incentives from curriculum tracking: cross-national and UK evidence

Kristian Koerselman*†‡§

February 16, 2011

Abstract

Curriculum tracking creates incentives before its start, and we should expect scores in tested subjects to be higher at that point. I find evidence from both UK and international data for sizable incentive effects. Incentive effects are important from a methodological perspective because they lead to downward bias in value-added estimates of the later age effect of tracking on achievement. They also invalidate placebo tests that work by regressing pre-tracking scores on tracking policies.

Keywords: *incentives, curriculum tracking, ability streaming, high-stakes testing, student achievement.*

JEL: *I21, I28, J08, J24*

*Swedish Institute for Social Research SOFI, Stockholm, Sweden

†Department of Economics, Abo Akademi University, Turku, Finland

‡Contact information at http://economistatwork.com

§I thank Tuomas Pekkarinen, Markus Jäntti, Ludger Woessmann, Heikki Kauppi, Jonas Lagerström, Sari Kerr, Roope Uusitalo, Elias Einiö, Sangjun Jeong, Fabian Pfeffer, Sharon Simonton and Anders Stenberg for their kind 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*.

1 Introduction

Curriculum tracking is the explicit separation of students into schools or classes based on observed past or expected future achievement. While it is uncommon to explicitly track on the primary level, and the norm is to do so on the tertiary, there are large differences in tracking policies on the secondary level. Since the Second World War, some countries have postponed tracking from the end of primary school to the end of middle school or even to the end of high school, while others have left their tracking policies unchanged (Benn and Chitty 1996, p. 7). This makes questions on the effects of tracking highly relevant. At the same time, the variation in tracking policies, both temporal and spatial, provides us with a means to identify its effects.

The literature has mainly focused on the long-term, net effect of tracking on educational achievement and wages, measuring outcomes after the end of compulsory education or later. While the effect of tracking on mean test scores may be positive (Kim et al. 2003, Galindo-Rueda and Vignoles 2004, Duflo et al. 2008) or negative (Hanushek and Woessmann 2006, Pekkarinen et al. 2009), there is a consensus among these authors that tracking also increases differences between students. Tracking probably reduces intergenerational mobility as well (Brunello and Checci 2007, Maurin and McNally 2008).

It is however also important to look at early-age effects of tracking policies on student outcomes. Specifically, tracking creates incentives before its start, amongst others for students to work harder and try to get into a higher track. The tracking point is thus a high-stakes moment for the student, whether the track choice is based on an explicit test or not. Bishop (1998) and Jacob (2005) find that high-stakes tests lead to higher student achievement. This is a subset of a more general literature which shows that students and teachers respond to incentives (Bishop 2006).

Waldinger (2006) mentions the possible existence of incentive effects, and in the model of Eisenkopf (2009), tracking makes educational signaling more efficient by shifting incentives to an earlier age. Galindo-Rueda and Vignoles

(2004) find incentive effects in UK data, but the main focus of their paper is on post-tracking outcomes.

I add to this literature by making a comprehensive empirical analysis of incentive effects of tracking. I show that incentive effects are well identified in UK data, and that a similar pattern can be found in international data.

Incentive effects have methodological implications. The existence of incentive effects makes value-added estimates of the later age effects of tracking (e.g. Hanushek and Woessmann 2006, cf. Todd and Wolpin 2003) misspecified. Pre-tracking test scores are not exogenous, but positively related to early tracking, leading to a downward bias in tracking estimates that use early test scores to control for unobservables.

A second implication that a positive relationship between pre-tracking scores and tracking policies cannot be used as an argument that there is selection in post-tracking regressions (e.g. Manning and Pischke 2006).

2 Incentives

In theory, the incentives from tracking may work in many ways. 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 the student a better peer group, which will increase his future achievement (Hoxby 2000, Ammermueller and Pischke 2006). Upper track attendance will also usually leave open the possibility to enter university 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.

The student may also substitute effort between subjects: from nontested subjects to tested ones. This is indeed the case in Jacob (2005), but not in Winters et al. (2008) who suggest that positive spillover effects from the tested subjects compensate for the crowding-out of nontested ones.

Teachers have an incentive to teach better as well as to substitute time and effort towards tested subjects. 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. Teachers do indeed change their behavior in expected ways in Jacob (2005).

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, 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.

Across countries, tracking policies may also affect the early curricula or teaching styles in a more institutionalized way. The whole educational system may have evolved towards stressing early achievement more. Of course, the direction of causality may also run the other way if early achievement oriented countries have refrained from delaying the tracking point (cf. Betts 2010).

To at least some degree, incentive effects cause students to do better at tests rather than learn more on an underlying level (cf. Klein et al. 2000, Jacob 2005). This is a problem if we want to use incentives to increase underlying achievement. For the methodological implications however, the measured scores are more relevant than underlying achievement. Incentive effects can lead to inflated test scores relative to long-term effects of underlying achievement, whether the disparity is caused by temporary bumps in underlying or in measured achievement.

3 UK evidence

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 LEAs change their policies, and the rate of change was initially limited. The hesitant Labour attitude was induced both by 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, Local Education Authorities 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. This lead the policy change to be 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 (2010). It follows all those born in Great Britain in the week of the 3rd of March 1958. The 1958 cohort turned 11 in 1969, when one part of them were 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. Merging the different sweeps, I have 6435 complete cases.

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 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, that is 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 normalize each test score distribution because their shapes are arbitrary and skewed, and contain little cardinal level information on underlying achievement (Koerselman 2011). 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 encode the school tracking status at age 11, T_s, as a school-level dummy indicating whether the school turned comprehensive before 1969, or after. I also select two groups of control variables, listed in Table 5 in the appendix. 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 choose not to linearize any of these variables and treat them all as categorical in order to capture as much variation as possible.

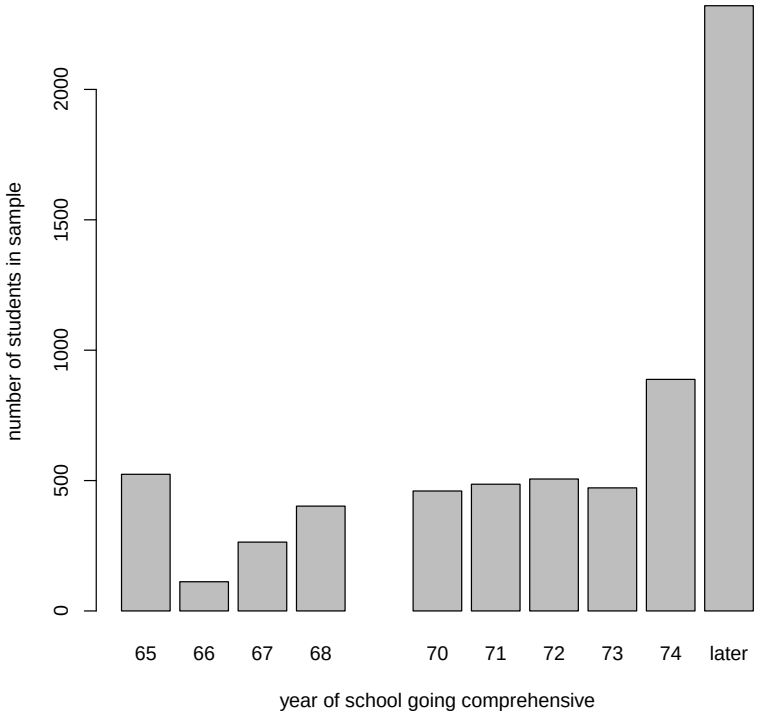

Figure 1: Year of school going comprehensive: number of students in the sample. The students in the sample all turned 11 in 1969, at which point they were split into tracks in the pre-reform system. Those entering secondary schools having reformed before 1969 (left) should be expected to have lower age 11 scores than those entering schools that reformed later (right).

Unfortunately for our purposes, reforms were not implemented at random. As can be seen from Table 5, right-wing, richer areas were underrepresented among the areas that moved to a comprehensive system first (Benn and Chitty 1996, ch. 1, Galindo-Rueda and Vignoles 2004), leading to a negative correlation between 1969 reform status and school inputs. A simple comparison of tracked and comprehensive areas or schools will therefore appear 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 background. Selection problems can however be expected to smaller than for later-age educational

analyzes because the primary school system is relatively homogeneous.

Additionally, there may be selection within and between regions due to non-compliance. Families with good students can 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 become the new lower track school, with the upper track 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.

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 levels. For example, in the first specification

$$y_i = \alpha + T_s\beta + \varepsilon_s + \varepsilon_i \tag{1}$$

individual achievement y_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.

$$y_i = \alpha + T_s\beta + A_i\gamma + \varepsilon_s + \varepsilon_i \tag{2}$$

$$y_i = \alpha + T_s\beta + A_i\gamma + X_i\delta + \varepsilon_s + \varepsilon_i \tag{3}$$

The results of these specifications can be seen from the Table 1. The first column shows the unadjusted relationship between age 11 scores and the tracking variable is 0.15 of a UK standard deviation. This is a sizable difference, but probably an overestimate of the causal effect since early reform areas were poorer on average.

Turning to column (2), we can see that the estimated effect indeed declines

Dependent variable: UK achievement age 11 (1969)						
specification	(1)	(2)	(3)	(4)	(5)	(6)
School not comprehensive at age 11 (T)	0.15	0.10	0.09	0.09	0.09	0.08
	0.04	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*
age 7 scores and ratings (A_i)		yes	yes	yes	yes	yes
additional controls (X_i)			yes	yes	yes	yes
number of students	6435	6435	6435	5109	6435	6435
grouping	schools	schools	schools	schools	LEAs	years
number of groups	616	616	616	528	156	10

Table 1: 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.

to 0.10. 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 changes very little between the specifications, and is now 0.09. This strongly suggests that age 7 test scores pick up most of the selection, and that even less selection will be left after the inclusion of X_i.

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 nonmovers only. This reduces the number of students from 6435 to 5109, and the number of schools from 616 to 528 (the sampling method causes individual schools to be represented by small numbers of students). As can be seen from column (4), the results are still unchanged at 0.09.

Next, I look at possible selection within areas by using the percentage 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. There are 156 LEAs in the sample. As can be seen from column (5) however, the point estimate is still unchanged. suggesting that within-LEA selection is not a problem given the controls available to us.

As an additional check, I group all schools together by reform year, and define

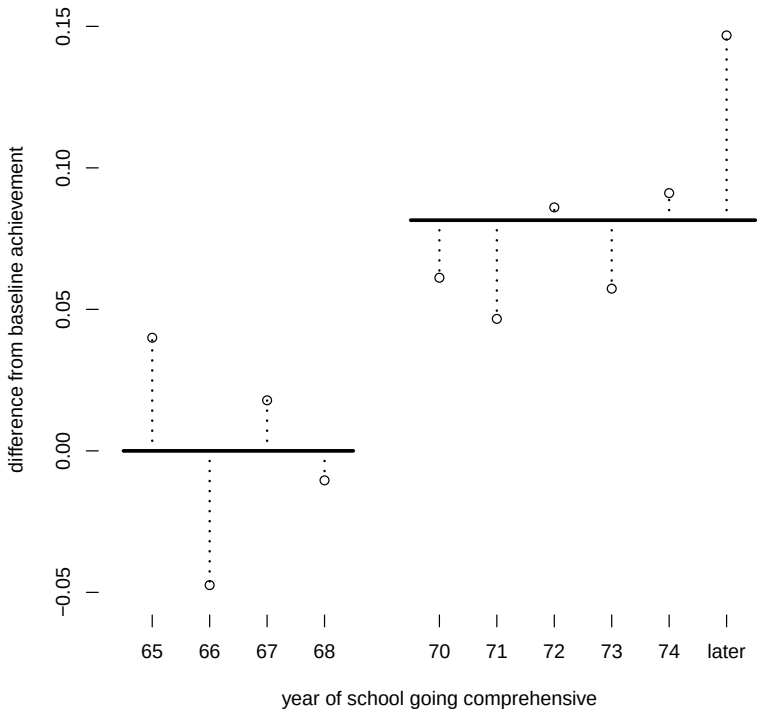

Figure 2: Secondary schools left of the divide turned comprehensive before the NCDS students could enter them. Achievement estimates from specification 6. Dotted lines indicate the year-level errors.

tracking as a year-level variable.

$$y_i = \alpha + \beta T_y + A_i\gamma + X_i\delta + \varepsilon_y + \varepsilon_i \tag{6}$$

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.

I also rerun specifications (1) and (2) with age 7 achievement as the dependent variable as a kind of placebo test under the assumption that incentive effects should be weaker the longer before the tracking point we measure

achievement. Unfortunately, we cannot control for early age scores when using them as the dependent variable. Still, as can be seen from Table 6 in the appendix, the estimated treatment effect is much smaller and not significantly different from zero for age 7 outcomes. This is 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. Incentive effects are not significantly different between boys and girls. I also add interactions on father's socioeconomic status to specification 2, but no monotonic pattern can be seen, and the uncertainty of the interactions is large. I have illustrated these results in Figure 3.

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.

4 International evidence

The International Association for the Evaluation of Educational Achievement administers various standardized tests in a large number of countries, which 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

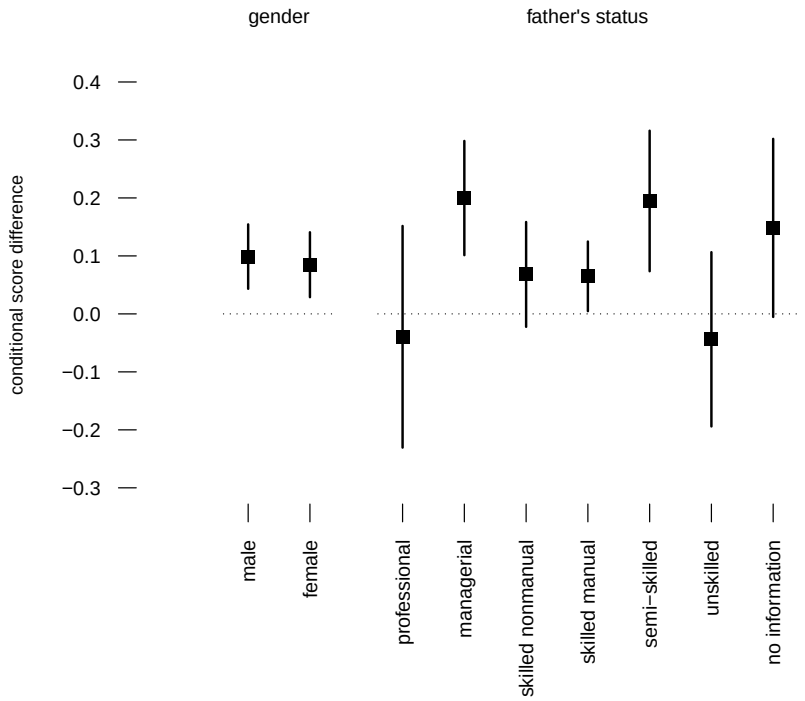

Figure 3: Estimated incentive effects for different subgroups. The gender-specific effect is conditional on all controls in specification 3 except height. The specific effect for different levels of socioeconomic status for the father is conditioned on all controls in specification 2. Bars indicate the 95% confidence interval. The size of the effect is not significantly different between boys and girls. No monotonic pattern can be found in the socio-economic background of the student.

the participating countries. I take the average of TIMSS mathematics and science scores to get a more general measure of achievement.

I take tracking information from the Eurybase database (Eurydice 2008), as well as from a variety of other sources. 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), and aids a comparison with their results. 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 14 or earlier.

variable	weighting by student μ	by student σ	by country μ	by country σ
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

Table 2: International data: descriptive statistics.

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 the 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 2. I have complete data on 1040596 students in 51 countries.

Like before, 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 given by

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

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_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 a matrix D_i indicating whether the score is a PIRLS or a TIMSS score.

$$y_i = \alpha + T_{cn}\beta + D_i\gamma + \varepsilon \tag{7}$$

Dependent variable: international early age achievement					
	(7)	(8)	(9)	(10)	(11)
tracking	0.26	0.11	0.22	0.23	0.25
	0.16	*0.13*	*0.07*	*0.06*	*0.07*
GDP		0.39	0.01	0.00	0.01
		0.07	*0.04*	*0.04*	*0.04*
expenditures		-0.08	0.03	0.02	0.03
		0.04	*0.02*	*0.02*	*0.02*
books at home				0.14	
				0.00	
tracking*books at home				-0.01	
				0.04	
female					0.05
					0.00
tracking*female					-0.05
					0.03
students	1040596	1040596	515788	515788	515788
countries	51	51	28	28	28

Table 3: International evidence for incentive effects; pooled multilevel regression based on international data. Standard errors in italics.

The results can be seen from column (7) in Table 3. Countries with early tracking clearly have higher score means, with the mean difference as large as 0.26 standard deviations of international student test scores.

There is no reason to assume that countries have adapted tracking policies at random, and the observed correlation may be mere selection. To make an attempt to control for this, I include real per capita GDP and educational expenditures in the next specification. Both variables are contained in the country level matrix C_{cn}.

$$y_i = \alpha + T_{cn}\beta + D_i\gamma + C_{cn}\delta + \varepsilon \tag{8}$$

The estimates from this specification can be seen from column 8. Estimated incentive effects are now smaller at 0.11 standard deviations.

There is probably still much unobserved heterogeneity left. Also, the tracking measure used is most relevant in a European context, as it classifies within-school tracking countries as late tracking (Betts 2010). For both reasons, I restrict the sample to the more homogeneous European Economic Area

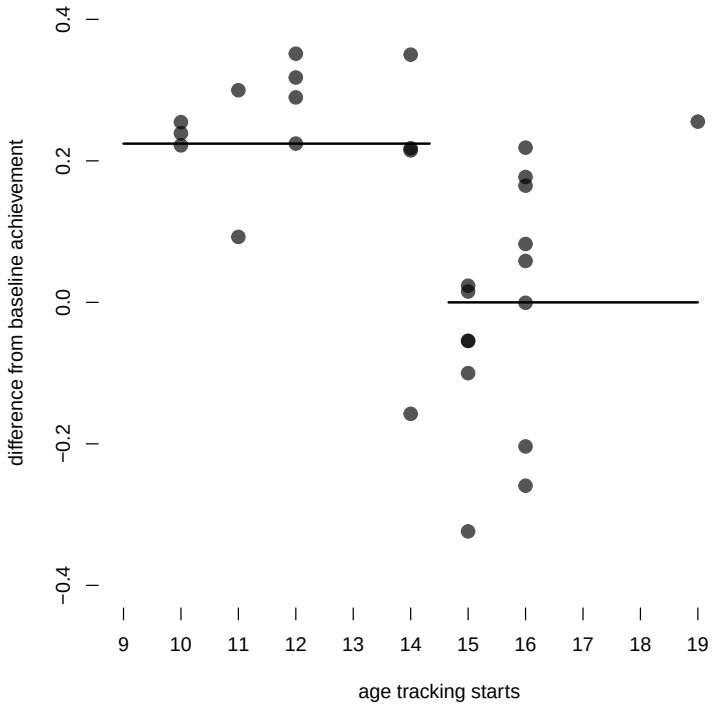

Figure 4: An illustration of the EEA estimate of incentive effects from specification (9). Early tracking countries have higher conditional early test scores. The solid line represents the estimate, circles indicate the country-level errors.

member countries, and rerun the previous specification.

The estimates from this specification can be seen from column (9). At 0.22, the effects are now much larger, but also much more precisely estimated. This is exactly what should be expected if the tracking variable has classical measurement error for non-EEA countries. Another indication that this is the better specification is that the estimated effect of educational expenditures now has the correct sign, even if it is still insignificant.

I have illustrated the estimate from specification (9) in Figure 4. As can be seen from the figure, a linear specification 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.

I try to estimate whether incentive 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. This variable is available for all four surveys.

Books at home are a good measure of parental background. The data are derived from a student questionnaire, and young children should be expected to report alternative measures of parental background such as educational attainment or exact occupation with considerable error. Books at home is also more easily compared internationally than education, occupation or income, it is a valid international proxy for family background (Schuetz 2008) and actually seem capture the reading culture driving the intergenerational transmission of educational attainment (Esping-Andersen 2004).

$$y_i = \alpha + T_{cn}\beta + D_i\gamma + C_{cn}\delta + B_i\theta + (B_i \cdot T_{cn})\kappa + \varepsilon \tag{10}$$

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

Results can be seen from column (10). 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.

$$y_i = \alpha + T_{cn}\beta + D_i\gamma + C_{cn}\delta + F_i\lambda + (F_i \cdot T_{cn})\mu + \varepsilon \tag{11}$$

Looking at column (11) of Table 3, we can see that the differences between boys and girls are small, and that the interaction is not significantly different from zero even though it is estimated at 0.05. Both the unclear differences in parental background and the insignificantly smaller incentive effects for girls mirror the UK findings.

Hanushek and Woessmann make a slightly different assessment of the track-

ing age variable, even if they are 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 (7) and (8), but makes no difference in the EEA sample of the later specifications.

All in all, international test score data provide us with some evidence for incentive effects of curriculum tracking. The tracking variable is highly significant at in the European sample, which is unusual for any analysis including so few country-level observations. Nevertheless, we should realize that cross-country comparisons are inherently sensitive to omitted variable bias.

5 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 seems to cause an incentive effect of 0.09 UK standard deviations. Within the European Economic Area, tracking is associated with 0.22 international standard deviations higher scores. These estimates are large, but of the same order of magnitude as 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 add a line of evidence to the UK results, where the effect seems well-identified.

The implications of incentive effects are twofold. On the one hand, they are of methodological importance. A causal effect of tracking extending to the age before it start implies that value added estimates (see e.g. Todd and Wolpin 2003) of the long-term effect of tracking are misspecified. Because pre-tracking scores are inflated in early tracking systems, a value-added spec-

category	authors	year	mean effect
comprehensive school reform, panel data	Pekkarinen et al.	2009	–
comprehensive school reform, cross-section	Kim et al.	2003	+
	Galindo-Rueda and Vignoles	2004	+
international cross-section	Hanushek and Woessmann	2006	–
experimental	Duflo et al.	2008	+

Table 4: Important studies of the mean effect of tracking.

ification which controls for omitted variables using pre-tracking scores will underestimate the later-age long-term effect of tracking.

If we accept the invalidity of value-added specifications, we can reconcile previous studies on the long-term effect of tracking. I have listed some current current papers on the mean effect of tracking in Table 4. The effect on the mean is negative in the panel data papers as well as in Hanushek and Woessmann.

We should not be surprised to find an apparent negative effect of tracking in studies of post war reforms such as Pekkarinen et al. The reforms simultaneously changed the tracking structure and upgraded the quality of education of those previously in the lower track. If a country with a modern vocational track such as Germany were to postpone it's tracking point today, the positive effects could be much smaller.

The other main study finding a negative effect is that of Hanushek and Woessmann. Hanushek and Woessmann however use a value-added specification, controlling for pre-tracking achievement. If one believes that tracking has incentive effects, this specification is invalid, and leads to downward biased estimates of the mean effect of tracking. They find an effect not significantly different from zero when omitting early scores.

The other authors all find a positive effect of tracking on mean scores. I thus conclude that a positive effect of tracking on mean test scores is the most consistent with the data. Of course, we should remember that the effects of tracking on inequality and intergenerational mobility are large, more certain and perhaps more important as well.

It should also be noted that 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. It is this very phenomenon which I describe as incentive effects. If we believe that measured incentive effects can be causal, we should therefore not reject the UK literature on these grounds.

References

Andreas Ammermueller and Joern-Steffen Pischke. Peer effects in European primary schools: evidence from PIRLS. ZEW discussion paper no. 06-027, 2006.

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.

John Bishop. The effect of curriculum-based external exit systems on student achievement. *Journal of Economic Education*, 29(2):171–182, 1998.

Giorgio Brunello and Daniele Checchi. Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, 52:781–861, 2007.

Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer effects and the impact of tracking: Evidence from a randomized evaluation in kenya. NBER Working Paper No. 14475, 2008.

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

G. Esping-Andersen. Untying the Gordian knot of social inheritance. *Research in social stratification and mobility*, 21:115–138, 2004.

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

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

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*, volume 625. Cambridge University Press Cambridge, 2007.

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

Caroline Hoxby. Peer effects in the classroom: learning from gender and race vatiation. NBER working paper no. 7867, August 2000.

International Association for the Evaluation of Educational Achievement IEA. Trends in International Mathematics and Science Study TIMSS. 1995.

International Association for the Evaluation of Educational Achievement IEA. Progress in International Reading Literacy Study PIRLS. 2001.

International Association for the Evaluation of Educational Achievement IEA. Trends in International Mathematics and Science Study TIMSS. 2003.

International Association for the Evaluation of Educational Achievement IEA. Progress in International Reading Literacy Study PIRLS. 2006.

B.A. 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.

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

Taejong Kim, Ju-Ho Lee, and Young Lee. Mixing versus sorting in schooling: evidence from the equalization policy in South Korea. KDI School Working Paper No. 03-07, 2003.

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

Kristian Koerselman. Bias from the use of mean-based methods on test scores. Swedish Institute for Social Research (SOFI) Working Paper 1/2011, 2011.

Alan Manning and Joern-Steffen Pischke. Comprehensive versus selective schooling in England and Wales: what do we know? NBER working paper no. 12176, April 2006.

E. Maurin and S. Mcnally. The Consequences of Ability Tracking for Future Outcomes and Social Mobility. *Centre for Economic Performance*, 2008.

National Child Development Study (NCDS). National Child Development Study 1958–. 2010.

Tuomas Pekkarinen, Roope Uusitalo, and Sari Kerr. School tracking and development of cognitive skills. VATT working paper 2, 2009.

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

Penn World Table 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. Schuetz, H.W. Ursprung, and L. Woessmann. Education policy and equality of opportunity. *Kyklos*, 61(2):279–308, 2008.

Petra Todd and Kenneth 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.

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

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

World Bank. EdStat Education Statistics. 2011.

Appendix

	overall		tracked	compr.
variable name	mean	sd	mean	mean
dependent variable y_i				
Achievement age 11	0.00	1.00	0.05	-0.18
early ability A_i				
Arithmetic score age 7	0.00	1.00	0.02	-0.07
Copying designs score age 7	0.00	1.00	0.01	-0.05
Drawing score age 7	0.00	1.00	0.01	-0.03
Reading score age 7	0.00	1.00	0.03	-0.13
Creativity rating age 7	0.00	1.00	0.02	-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
Additional controls X_i				
Female	0.49		0.49	0.50
Height age 11				
1st quintile group	0.19		0.18	0.21
1st quintile group	0.19		0.19	0.17
2nd quintile group	0.18		0.19	0.18
3rd quintile group	0.19		0.19	0.18
4th quintile group	0.19		0.19	0.18
5th quintile group	0.07		0.07	0.08
Father figure				
natural father	0.92		0.92	0.90
other	0.05		0.05	0.06
no information	0.03		0.03	0.04
Attended nursery				
public	0.02		0.02	0.03
private	0.04		0.04	0.02
other preschool	0.03		0.03	0.04
did not attend or no information	0.91		0.91	0.92
Father reads to child				
often	0.33		0.34	0.31
occasionally	0.33		0.33	0.34
hardly ever	0.26		0.26	0.27
no information	0.07		0.07	0.07
Mother reads to child				
often	0.46		0.47	0.43
occasionally	0.34		0.33	0.36
hardly ever	0.15		0.15	0.16
no information	0.04		0.04	0.05
Socio-economic status father				

continued on next page

continued from previous page

	overall		tracked	compr.
variable name	mean	sd	mean	mean
professional	0.04		0.05	0.03
manegerial/technical	0.16		0.17	0.13
skilled nonmanual	0.09		0.09	0.09
skilled manual	0.43		0.42	0.45
semi-skilled	0.16		0.16	0.17
unskilled	0.05		0.05	0.06
no information	0.06		0.06	0.06
Father's education ISCED				
5	0.03		0.03	0.02
3	0.17		0.17	0.15
2	0.54		0.54	0.52
1	0.01		0.01	0.02
no information	0.25		0.24	0.29
Mother's education ISCED				
5	0.02		0.02	0.01
3	0.20		0.20	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 reads books				
often	0.47		0.48	0.42
occasionally	0.20		0.19	0.23
hardly ever	0.27		0.26	0.27
no information	0.07		0.07	0.08
Mother reads books				
often	0.32		0.33	0.29
occasionally	0.21		0.21	0.21
hardly ever	0.42		0.41	0.44
no information	0.05		0.05	0.05
Accomodation type				
house	0.86		0.86	0.84
flat	0.07		0.07	0.07
rooms	0.01		0.01	0.02
no information or other	0.05		0.00	0.00
Father born				
British Isles	0.92		0.93	0.91
Eire or Ulster	0.03		0.04	0.03
other	0.04		0.04	0.06
Mother born				
British Isles	0.93		0.94	0.91
Eire or Ulster	0.03		0.03	0.03
other	0.04		0.03	0.06
Poor at English age 7				
no	0.97		0.98	0.96
somewhat	0.01		0.01	0.02
certainly	0.00		0.00	0.01

continued on next page

continued from previous page

	overall		tracked	compr.
variable name	mean	sd	mean	mean
no information	0.01		0.01	0.02
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
Number of students	6435		5133	1302
Number of schools	616		450	166

Table 5: NCDS: student-weighted descriptive statistics.

dependent variable	age 11		age 7	
specification	(1)	(2)	(3)	(4)
School not comprehensive at age 11 (T)	0.15	0.13	0.05	0.03
	0.04	*0.03*	*0.04*	*0.04*
controls (X_i)		yes		yes
number of students	6435	6435	6435	6435
grouping	schools	schools	schools	schools
number of groups	616	616	616	616

Table 6: Placebo test for UK incentive effects using early age scores.