Incentives from curriculum tracking

Kristian Koerselman[a]

[a] *Helsinki Center of Economic Research, PO Box 17, 00014 University of Helsinki, Finland.*

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 find robust evidence for incentive effects of tracking in the UK based on the UK comprehensive school reform. Results from the Swedish comprehensive school reform are inconclusive. Internationally, I find a large and widening test score gap between early and late tracking countries. 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
JEL: I21, I28, J08, J24

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 large and robust test score gap between early late tracking countries that widens between third and fourth grade. Estimates based on a Swedish school reform are inconclusive.

Email address: koerselman@economistatwork.com (Kristian Koerselman)
URL: http://economistatwork.com (Kristian Koerselman)

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, Schneider, Senkbeil and Carstensen, 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.

There are many mechanisms through which tracking can increase early test scores. There is a direct incentive for students at the margin of upper track entry to try to get into the upper track. Attending the higher track will give students a better peer group, will usually also 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. Teachers can also be incentivized because the track placement of their students is easily observable, making it easier for for principals to reward and punish teacher effort as well as easier for parents to choose higher achieving schools for their children.

Students and teachers may substitute effort 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, Greene, and Trivitt, 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).

All of these mechanisms suggest that we should expect early test scores to be higher under an early tracking regime relative to a late tracking regime. The effects should be particularly large for students close to the margin of upper track attendance. Very good or very poor students may not feel directly incentivized because they cannot influence their track placement in any case. The effects on these students do however not have to be zero. There may be positive or negative peer effects from other students, and all students will be influenced by changes to the curriculum or teaching style.

It seems reasonable to think that test score increases do not (only) occur instantaneously at the tracking point itself, but that there is a period of time leading up to the tracking point during which students' test scores grow at an accelerated pace. It is less clear whether we should expect students in the late tracking regime to catch up during middle school or not. Test scores certainly converge after the start of early tracking in Hanushek and Woessmann (2006), but it is hard to tell whether this is because of to incentive-induced catch-up growth among late tracking countries or because comprehensive schools are genuinely more efficient.

2. UK evidence for incentive effects

Since the Second World War, the larger part of the UK has gradually gone from a tracked to a comprehensive school system. In the old system (Figure 1, top half), students took the co-called *Eleven Plus* achievement test around age 11. Those who did well enough on the Eleven Plus were allowed to enter a upper track *grammar school*. Grammar school students could acquire an Ordinary Level General Certificate of Education or O-level at age 16, and an Advanced Level General Certificate of Education or A-level at age 18, after which they could enter higher education. Students who failed to qualify for grammar school usually entered a vocational *secondary modern*. Secondary modern students could either acquire a Certificate of Secondary Education or CSE at age 16, or leave the education system one year earlier. The highest grade on the CSE was regarded as equivalent to an O-level.

Because the upper track was limited in size, and because the admittance procedure was noisy, not even top students could be certain of a place in the upper track. This can also be seen from Figure 2, where I have plotted the 10th, 50th and 90th percentiles of the predicted probability of upper track attendance based on age 7 test scores, parental educational attainment and socio-economic status and other background variables. The median predicted probability of entering the upper track was less than 10%, and even in the 10th decile group of age 7 test scores, the median predicted probability of upper track attendance was only marginally higher than 50%.

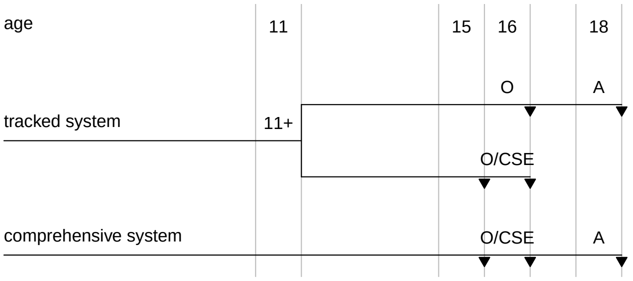

Figure 1: The main British secondary school systems around 1969.

In the tracked system (top), an age 11 test determined access to the upper track. In the comprehensive system (bottom), all students attended the same middle school. Triangles indicate common exits from secondary education: at the end of compulsory education but before taking the O-level or CSE examinations at the end of middle school; after completing middle school; after taking the A-level examinations at the end of high school.

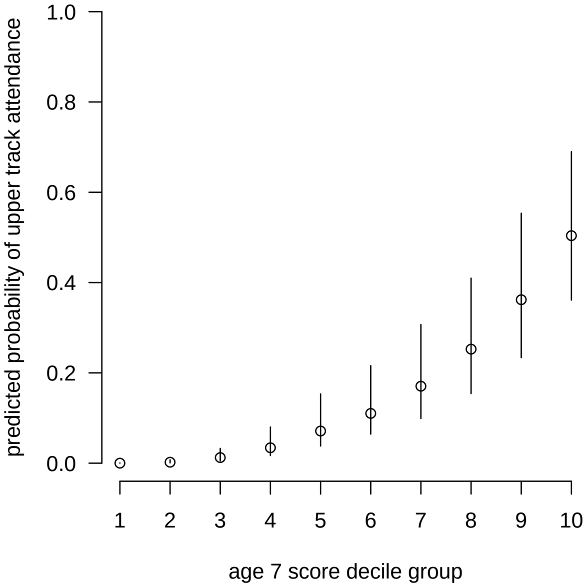

Figure 2: Limited capacity in upper tracks.

The figure shows the median predicted probability of upper track attendance (circles) as well as its 10th–90th percentage range (lines) by age 7 test score decile group. Though upper track attendance was unlikely for low achievers, not even high achievers could be certain of a place in the upper track. Data: NCDS (University of London 2008).

In 1964, the Labour government entered the elections with a promise to abolish the tracked educational system, and wanted to impose the new comprehensive system "as rapid as possible." Once in government however, the Labour government "requested" rather than demanded that Local Education Authorities (LEAs) change their policies.

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 schooling was 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, Fogelman, Crook and Reeder, 1996, ch. 2)

The new comprehensive schools aimed to make available to all children "all that is valuable in grammar school education" (Government Circular 10/65, 1965). Even though the different comprehensive schools were organized in different ways, the key characteristic of the new schools was that selection did not take place at age 11, but rather through voluntary exit at the compulsory schooling age of 15 or otherwise at the CSE or O-level examination. I have illustrated this in the bottom half of Figure 1.

When using comprehensive school reforms to identify the effects of tracking, it is important that the reform did not include other simultaneous changes that also affect outcomes. In this particular case, we are interested in the effects on test scores at the end of primary school. To the best of my knowledge, no simultaneous changes were made to the British primary school system in the years leading up to the reform. This stands in contrast with the Swedish comprehensive school reform covered below, where simultaneous curricular and other changes were made to primary schools.

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.

Table 1: NCDS sample sizes.

	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

Note: The main reason for missing tracking information is students attending private schools.

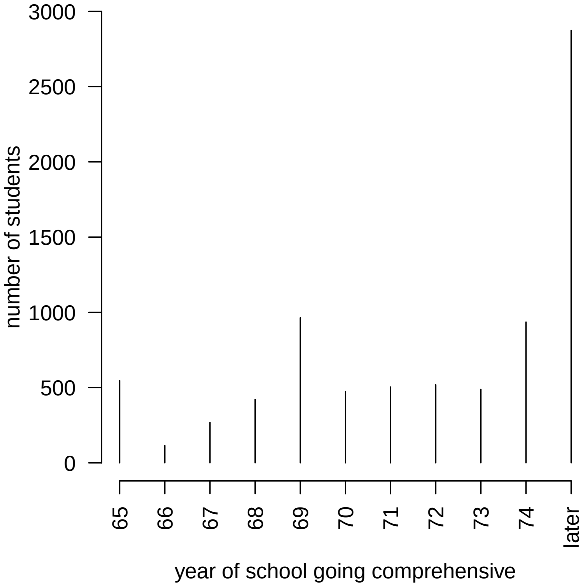

Figure 3: Number of students in sample by reform year.

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 comprehensive secondary schools (reform year before 1969) should be expected to have lower age 11 scores than those entering schools that reformed after 1969.

A small number of private schools indicate in the survey that they are comprehensive. 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 3. The students on the left side of the figure entered a secondary school that had reformed before 1969, and could be sure of its comprehensive status. Those on the right side entered a school that reformed only after 1969, and so faced a tracked system when they did the test. 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. I encode tracking status at age 11 (T) as a dummy indicating whether the student's school turned comprehensive before 1969, or after.

There are multiple measures of age 11 achievement in the data that can be used as outcome variables: 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.

Using test scores as outcome variables is problematic in various ways. Test scores usually only carry ordinal information. This means that both the absolute and relative score distances between students in different parts of the distribution are arbitrary (cf. Koerselman, 2011). In the case of the NCDS test scores, some of the test score distributions are strongly negatively skewed. There is however no reason to assume that the distribution of achievement has such skew, and compared to a symmetric distribution using untransformed test scores would have the effect of weighting the outcomes of low achieving students heavier.

Instead, I assume that student achievement is an unidimensional, normally distributed latent trait, and try to extract it. First, I convert all test score distributions to z-scores. Then, I take the first principal component of the transformed scores to end up with a measure of general achievement. This process also reduces measurement error from any of the specific tests.

Even though test scores now follow a symmetric distribution, the units the scores are expressed in are still meaningless. I therefore express outcomes in standard deviations. This however means that measurement error in the test score will lead to bias in the estimated treatment effects as expressed in standard

Table 2: NCDS, means of selected variables

	tracked areas	comprehensive areas
father highly educated	0.20	0.17
father high SES	0.22	0.17
father reads to child every week (age 7)	0.34	0.31
child height below first quintile (age 11)	0.18	0.21
arithmetic score (age 7)	0.01	-0.06

Notes: A father is counted as highly educated if his ISCED level is 3 or higher. A father is counted as high SES if his social class is professional or manegerial/technical. The first four variables have been constructed for illustrative purposes. The regressions include indicators for each level of the original variables.

deviations of the latent trait. Test noise adds to test score variance, and the noisier the test is, the more we will overestimate the width of the latent trait distribution. Because we divide our estimates by the standard deviation of the distribution, the estimates will be biased towards zero if test scores are noisy. Intuitively, the noisier a test is the less treatment will be able to change standardized test outcomes.

To solve this problem, I calculate the reliability ratio of the age 11 principal component 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. Estimated noise is however low for the principal component with reliability ratios close to unity. The difference between this method and simply reporting effect sizes of the principal component is therefore small in practice.

I select two groups of control variables. The first group A_i consists of the z-scores of age 7 tests 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. It consists of parental social status and a measure of their wealth, parental educational attainment and reading habits, measures of parental involvement in the child's education at age 7, parental region of birth, the student's gender as well as the student's within-gender height quintile group at age 11 – a proxy for health.[1]

The variables in X_i are meant are meant to as fully as possible catch the determinants of age 11 achievement. For example, students may do worse on the age 11 tests if they are of poor health, if they come from a lower class, lower educated family without a reading culture at home, if their parents are uninterested in their education, or if their parents are migrants. All these variables are

[1] NCDS variables n1687, n199, n186, n187, n43, n44, n1434, n1436, n622 and dvht11. I refer to the NCDS documentation for a more detailed description of these variables. In addition to these, I use the ISCED level of parental education as calculated by Sharon Simonton.

qualitative, and to capture as much variation as possible X_i contains dummy variables for each of the values that they take.

Individual and LEA identifiers are available directly from the data, but in addition I construct school identifiers in such a way that students are treated as attending the same school if their school was located in the same LEA and reformed in the same year. This reduces problems with the reform-induced restructuring of LEA schools.

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 schools, but I also estimate models where the higher level is the LEA or the reform year instead.

While multilevel models differ conceptually from methods based on least squares, the estimates of the incentive effects as well as the associated standard errors are always based on the higher level and are close to those obtained from OLS on data aggregated to the higher level.

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). The consequences can also be seen from Table 2. Students in reform areas had less educated fathers, with lower socioeconomic status, experienced less parental involvement in their education and were shorter on average, an indication of relatively poor health. Unsurprisingly, they also had lower test scores at age 7.

Because of the nonrandomness of the reform, a simple comparison of tracked and comprehensive areas or schools s

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

is 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.

I claim that controlling for age 7 scores A_i

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

is enough to remove almost all of the bias due to the the nonrandomness of the reform. A natural way to check this is to add individual background variables as well

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

and see if the estimate of β changes much. If it does not, this suggests that controlling for age 7 scores is indeed removing most of the selection.

There is a second kind of selection problem. Even if the reforms would have been carried out at random, students may not have complied with the treatment assigned to their school by moving. Families with good students have an incentive to 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 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.

In specification (4), I restrict the sample to students who 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 because the sampling method causes individual schools to be represented by small numbers of students.

If between-region selection is a major problem, the estimate of incentive effects should be considerably lower in this specification, since it excludes the presumably higher scoring students that moved to tracked areas and the presumably lower scoring students that moved to comprehensive areas.

In specification (5), I group by Local Education Authority: the policy-setting authority of which there are 167 in the sample. I use the proportion of tracked students within the LEA as a measure of tracking. If observed differences between tracked and comprehensive students are due to poor students sorting into comprehensive schools and good students into upper track schools within mixed LEAs, we, we should expect the estimate of incentive effects to be considerably lower in this specification as well. Since tracking is an LEA-level variable in specification (5), this specification also addresses concerns that the standard errors reported for the other specifications may be too small since they are calculated on schools rather than on LEAs.

Incentive effects can both be direct through changed student incentives or indirect though peer effects and changed teacher behavior. We should expect lower direct incentive effects among those who have little hope of entering the upper track in any case as well as among those who could be certain of a place in the upper track. As can be seen from Figure 2, the upper track was extremely selective. Only very few students could be certain of a place in the upper track while the median student had an estimated probability of upper track attendance of slightly less than 10%.

In order to try to separate direct from indirect effects, I create an additional variable indicating whether the student has a predicted probability of upper track attendance of less than 10%. I regard these students as not competitive for the upper track. These students should thus mostly be exposed to indirect incentive effects. In specification (6), I interact this variable with T to see how much of the incentive effects are direct.

Additionally, I try to look for evidence of heterogeneous effects with regards to the timing of the reform. Those who entered a comprehensive secondary school could be certain of its status, but those entering a school that still had to reform will have experienced stronger incentives the further away they perceived the reform to be. We should expect students entering tracked schools not only

Table 3: Incentive effects in the UK.

specification	(1)	(2)	(3)	(4)	(5)	(6)
tracking (T)	0.148	0.098	0.095	0.098	0.102	0.169
	0.044	*0.023*	*0.022*	*0.023*	*0.034*	*0.029*
$T\times$not competitive						-0.138
						0.036
ability (A_i)		yes	yes	yes	yes	yes
controls (X_i)			yes	yes	yes	yes
students	7150	7150	7150	5634	7150	7150
groups	645	645	645	556	167	645
grouping level	school	school	school	school	LEA	school

Notes: Dependent variable: UK achievement age 11 (1969). Standard errors are calculated on the (higher) grouping level, and are shown in italics.

to have higher test scores than those entering comprehensive schools, but also to have higher test scores the further the reform lay in the future.

Results from the first six specifications can be seen from Table 3. Column (1) shows that the unadjusted relationship between the tracking variable and age 11 scores is 0.148 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 0.098 standard deviations when we control for the nonrandom nature of the tracking reforms using age 7 scores. Controlling by age 7 scores only may not be enough to remove all selection, and I therefore add the wide range of background variables described above to the next specification. The result can be seen from column (3). The estimate only decreases by 0.003 standard deviations. This strongly suggests that age 7 test scores pick up most of the selection, and that the nonrandom implementation of the reform is not an obstacle to identification of its effects on early test scores.

In column (4) we see the regression results using nonmovers only. The sample size is now smaller, reducing precision slightly. Importantly, the estimate of the effect of tracking is now larger. This suggests that conditional on covariates, selective moving between LEAs is not a problem for identification.

In column (5) we see the LEA level regression results. Again, estimates are larger than under the baseline specification (3). This suggests that selective moving within the LEA is not an obstacle to identification either. A closer look at the LEA-level errors (available from the author) shows that incentive effects are almost linear in the proportion of the LEA that is tracked, also suggesting that within-LEA selection is not a large problem.

Because students are grouped at a higher level in specification (5) than in specifications (1) through (4), the standard errors are somewhat larger. However, estimates are still significant at the 5% level, and the use of this higher grouping level does not change any qualitative conclusions.

Column (6) shows the results of adding an interaction term between tracking and the student not being competitive for the upper track. The estimates show that the incentive effect on competitive students is about 0.17 standard deviations, and that the effect on those that are not competitive is about 0.03. This suggests that direct incentive effects are more important than net indirect effects through teachers and peers.

The results in column (6) are also an additional indication that the observed effects are due to the reform rather than selection. If the results were due to selection, we would expect the effects on uncompetitive students to be much larger.

In Figure 4, I have illustrated estimated incentive effects by year. The estimated difference between tracked and comprehensive regions, as indicated by the difference between the two solid lines, is now slightly lower because the results are weighted by year rather than by school. As expected, an increasing pattern is visible in the right half of the figure, with test scores being higher for students whose schools reformed later. This suggests that incentive effects were not as large for those that knew that could expect that even if they failed their age 11 test, they would still have ended up in a comprehensive school by the time they exited secondary education.

Summarizing, the identification of incentive effects from British data looks credible. There were no substantial changes made to the primary school system at the time, implying that the measured reform effect plausibly represents changed incentives. The biggest other threats to identification are the nonrandom 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 I add a wide range of background variables as controls, lending credibility to the identification strategy. Neither excluding movers nor using LEA-level tracking variables changes the point estimate much. Incentive effects occur almost exclusively among students who stood a reasonable chance of entering the upper track at all. This adds further support to hypothesis that the observed differences are indeed due to incentive effects.

3. 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)

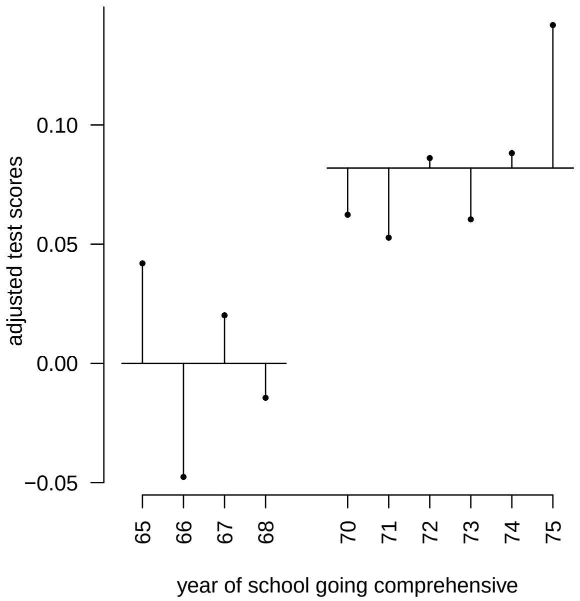

Figure 4: Incentive effects in the UK.

The figure shows regression-adjusted test score differences by reform year. Secondary schools left of the divide turned comprehensive before the NCDS students could enter them, and did not provide the incentives that the tracked schools on the right side provided them with. Multilevel regression by reform year. The horizontal lines represent the year-level estimate ($p<0.05$), dots indicate year-level errors.

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. For example, 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)

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 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 a majority of 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 4, this decreases to 19946 students in 1013 municipalities for which I have information on IQ, and further to 17427

Table 4: Number of observations in the UGU 1948 and 1953 cohorts at age 12.

	students	municipalities		
		total	tracked	comprehensive
all	21877	1020		
with IQ	19946	1013		
with math score	17427	1005		
in 1960	9290	946	801	145
in 1965	8137	930	313	617

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 were not in 6th grade at age 12 and of those that had transferred to an upper track school at an earlier age. There is thus a direct link between the comprehensive school reform and missingness on the mathematics test, at the very least because the comprehensive school reform abolished the upper track schools and subjected those students to the primary school math test. I will look at the effects of excluding this group further below.

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, and so a sizable proportion of municipalities changed tracking policies between those years. To illustrate the sensitivity of the results, I also use an alternative measure which Holmlund (2007) has collected from administrative sources. The correlation between the two measures is slightly above 90%.

I estimate variations of a fixed effects model

$$y_i = \alpha + T_i\beta + M_i\gamma + C_i\delta + X_i\zeta + Z_i\theta + \varepsilon_i \tag{7}$$

where y_i is a test score outcome, T_i municipal tracking status, M_i a matrix of municipality indicators, C_i a matrix of cohort indicators, X_i a matrix with municipality×cohort background variables, Z_i a matrix of individual background variables, and ε_i the error term. I weight individual observations with the inverse of the number of observations per municipality×cohort, and use standard errors clustered on the municipality level.

I have listed estimation results in Table 5. As can be seen from column (7), the effect of the reform on IQ scores seems to have been slightly negative, both for the tracking measure based on UGU and for the tracking measure based on administrative sources. In both cases, the null of no effect on IQ fits comfortably within the confidence intervals.

It is possible that we do not see a reform effect on IQ because IQ scores are harder to change through effort than other measures of achievement. I therefore also estimate effects on math scores. Columns (8) through (10) show results for the subsample for whom math scores are known. As can be seen from column (8), the reform effect on IQ is more negative in this sample, probably due to

Table 5: The effects of the Swedish comprehensive school reform on early test scores.

Dependent variable:	IQ (7)	IQ (8)	math (9)	math (10)
Tracking information from UGU				
tracking	-0.02	-0.07	-0.05	0.00
	0.05	*0.05*	*0.05*	*0.04*
ability controls				yes
observations	19946	17427	17427	17427
municipalities	1013	1005	1005	1005
Tracking information from administrative sources				
tracking	-0.01	-0.04	0.01	0.05
	0.05	*0.05*	*0.05*	*0.04*
ability controls				yes
observations	19916	17402	17402	17402
municipalities	1009	1002	1002	1002

Notes: All specifications include municipal fixed effects, cohort fixed effects and municipal and individual background characteristics. Municipality level clustered standard errors in italics.

selectively missing observations.

Column (9) shows the apparent effect on math scores in the subsample, but is it unclear what part selectivity plays in this estimate. Under the assumption that the true reform effect on IQ is zero, we can try to control for selectivity by including the three IQ subscores in the regression. I do this in the last column. The estimate of the reform effect on math scores is now zero for the UGU based tracking measure, and moderately positive for the tracking measure based on administrative sources.

In the end, it is hard to draw conclusions on incentive effects of tracking based on the Swedish comprehensive school reform. None of the estimates are significantly different from zero. At the same time, confidence intervals are wide enough to include effects of the size estimated in the UK.

Moreover, the reform changed many aspects of education simultaneously, and what we measure are the combined effects of multiple mechanisms. The reform changed the pre-test curriculum and pre-test teaching styles, it lowered the cost of continued education and changed the amount of compulsory education. It is possible that what we are measuring is a positive incentive effect of tracking canceled out by other aspects of the reform. In this respect, the British reform was a much cleaner policy experiment than the Swedish one.

4. A cross-country comparison

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

Both surveys aim to test a representative sample of the population of fourth graders in the participating countries, with the exception of TIMSS 1995 which also tests third graders. In the main specification, I remove the third graders from the sample in order to make the test scores more comparable across surveys.

I make no attempts to estimate reliability ratios 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. 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. To make it easier to interpret the estimates, I have listed the means and standard deviations of these variables in cTable 6.

I limit the sample to countries that are members 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 does not capture the within-school tracking and other, less explicit forms of stratification present in other parts of the world (Betts, 2010). The sample size for the baseline regressions is 471638 students in 28 countries.

As in the analysis of the UK school reform, I estimate a multilevel model to take into account the correlated outcomes of individuals that 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. Standard errors are calculated at the country level.

Table 6: Cross-country data: means and standard deviations.

	weighting			
	by student		by country	
variable	μ	σ	μ	σ
test score	0.00	1.00	-0.14	1.00
per capita GDP ('0 000 1995 USD)	0.32	0.47	0.32	0.47
educational expenditures (%GDP)	1.76	0.54	1.54	0.70
students				471638
countries				28

The first specification gives the raw relationship between individual scores $y_{cn,s,cl,i}$, and the country-level tracking regime T_{cn}. 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 7. Countries with early tracking clearly have higher score means, with the mean difference as large as 0.35 standard deviations of EEA student test scores.

There is no reason to assume that the estimated effect is not due to some third factor. I 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). Differences in economic development are not behind the correlation between tracking and early test scores. Estimated incentive are now even slightly higher at 0.38 standard deviations.

I have illustrated the estimate from specification (12) in Figure 5. Test scores are clearly decreasing in the number of years the tracking point lies ahead of grade 4. 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.

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 leaves my results almost unchanged.

Because TIMSS 1995 tests at both grade three and four, we can see if the estimates are larger in fourth than in third grade. To do this, I include an interaction between tracking policies and grade in a sample limited to TIMSS 1995. As can be seen from column (13), the test score gap between early and

Table 7: The relationship between tracking and early test scores across countries.

Dependent variable: early age achievement	(11)	(12)	(13)
tracking (T)	0.35	0.38	0.44
	0.10	*0.11*	*0.19*
GDP		-0.01	-0.04
		0.07	*0.16*
expenditures		0.03	0.05
		0.04	*0.07*
grade four			0.82
			0.03
$T\times$grade four			0.10
			0.05
students	471638	471638	151108
countries	28	28	14

Notes: GDP is per capita, and expressed in tens of thousands of 1995 US dollars. Educational expenditures are expressed as a percentage of GDP. Country level standard errors in italics.

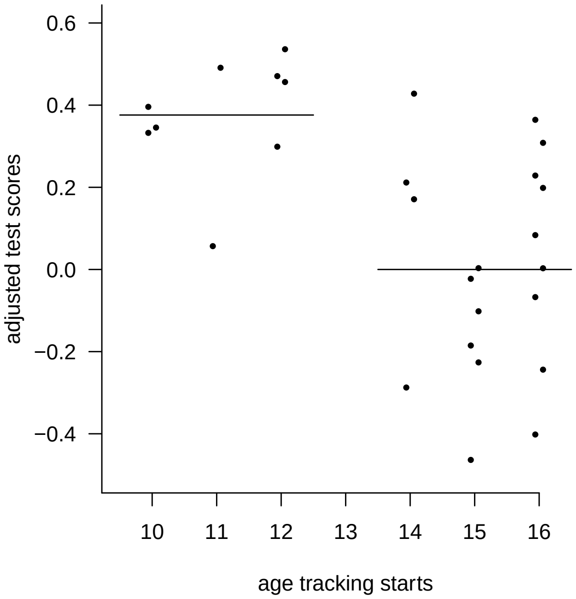

Figure 5: International test scores by age of tracking.

The figure is an illustration of specification (12). Early tracking countries have higher regression-adjusted early test scores. The solid lines represent the estimate of β, dots indicate the country-level errors. The horizontal axis has been jittered slightly to improve visibility.

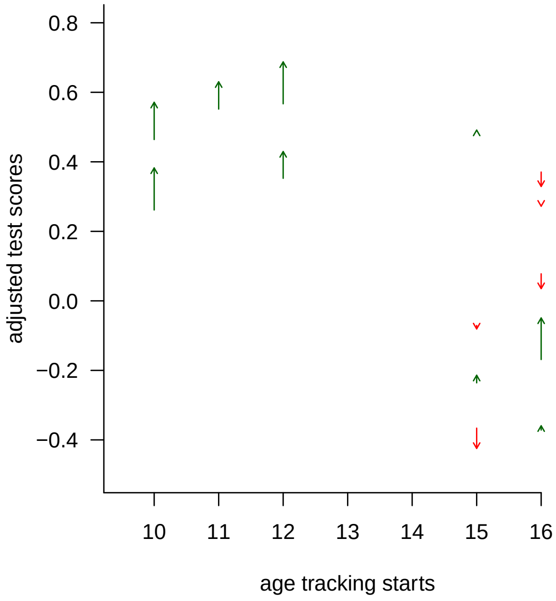

Figure 6: A widening gap in international test scores.

The gap between early and late tracking countries' regression-adjusted test scores increases between grades three and four. Arrows are drawn between countries' third and fourth grade regression-adjusted means so that the size and direction of each arrow indicates that country's score change between grades three and four.

late tracking countries is significantly diverging between grades three and four. As can also be seen from Figure 6, all five early tracking countries on the left side of the figure increase their relative performance between the two grades.

All in all, it is clear that there is a robust and strong gap in early test scores between early and late tracking countries. This gap is widening between grades three and four, and as we know from Hanushek and Woessmann, closing between grades four and eight or nine.

Even if these measured international differences are unlikely to to reflect an unidirectional causal link between tracking and early test scores only, they illustrate a remarkable but poorly understood pattern. My analysis shows that factors related to economic development are unlikely to be behind the gap, while the increase in the test score gap between grades three and four is suggestive of incentive effects as at least a partial explanation. The fact that we do not know what other factors contribute to the widening and closing of the test score gap only underlines that we should take great care in interpreting and using international test score data.

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 causes an average incentive effect of around 0.10 UK standard deviations. This effect is almost entirely driven by those who have a reasonable probability of entering the upper track, for whom the estimated effect is 0.17 standard deviations.

Within the European Economic Area, tracking is associated with 0.38 standard deviations higher scores. Both estimates are large, but the UK estimate is smaller than the 0.2–0.3 Jacob (2005) finds for a high-stakes test, while the international estimate is probably including other factors in addition to the causal effect of tracking on early test scores.

Estimates based on the Swedish comprehensive school reform are centered around zero. This can be explained by the fact that the Swedish reform consisted of many simultaneous policy changes, also to primary schools. At the same time, precision is too low to rule out sizable positive or negative effects of the Swedish comprehensive school reform on early test scores.

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 the 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.

In international data, the test score gap between early and late tracking countries increases between grades three and four, then closes between grades four and eight or nine. Hanushek and Woessmann (2006) argue that the comprehensive middle schools catch up to tracked ones because they are more efficient, but this need not be true if what we observe is actually the fading out of differential early incentive effects.

This point holds even if the widening and subsequent closing of the test score gap is due to other factors correlated with, but not caused by tracking. For example, if the countries that have gone through comprehensive school reforms are also the countries that concentrate on social and other noncognitive skills in primary school but catch up on hard skills in middle school, value-added estimates will still fail to recover the causal efficiency effects of comprehensive middle schools.

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 however 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.

Incentive effects of 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. 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 if comprehensive students do indeed catch up, its effects on later age outcomes may not be very large. It is not even clear whether increased early test scores reflect improved learning rather than improved test-taking skills (cf. Klein, Hamilton, McCaffrey and Stecher, 2000; Jacob, 2005; Almlund, Duckworth, Heckman and Kautz, 2011, section 5.6). Increased incentives of tracking can also have more direct negative effects on intrinsic motivation and well-being (cf. Juerges, Schneider, Senkbeil and Carstensen, 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 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.*

References

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

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

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

Bishop, J. (2006). 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.

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

Eurydice information network on education in Europe (2008). Eurybase database on education systems in Europe. Http://www.eurydice.org.

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

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

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

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

Härnqvist, K. (2000). 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*, . Distributer: Swedish National Data Service (SND).

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

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

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

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

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

Juerges, H., Schneider, K., Senkbeil, M., & Carstensen, C. (2012). Assessment drives learning. the effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, *31*, 56–65.

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

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

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

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

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

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

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

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

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

PWT (2006). 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.

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

Rindermann, H. (2007). 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*, 667–706.

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

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

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

University of London. Institute of Education. Centre for Longitudinal Studies (2008). 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.

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

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

World Bank (2011). EdStat Education Statistics.