ifo Working Papers

Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries

Eric A. Hanushek

Ludger Wößmann

Ifo Working Paper No. 1

January 2005

Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence across Countries*

Abstract

Even though some countries track students into differing-ability schools by age 10, others keep their entire secondary-school system comprehensive. To estimate the effects of such institutional differences in the face of country heterogeneity, we employ an international differences-in-differences approach. We identify tracking effects by comparing differences in outcome between primary and secondary school across tracked and non-tracked systems. Six international student assessments provide eight pairs of achievement contrasts for between 18 and 26 cross-country comparisons. The results suggest that early tracking increases educational inequality. While less clear, there is also a tendency for early tracking to reduce mean performance. Therefore, there does not appear to be any equity-efficiency trade-off.

JEL Code: I2.
Keywords: Tracking, streaming, ability grouping, selectivity, comprehensive school system, educational performance, inequality, international student achievement test, TIMSS, PISA, PIRLS.

Eric A. Hanushek
Hoover Institution
Stanford University
CESifo and NBER
Stanford, CA 94305-6010, United States
Phone: (+1) 650 / 736-0942
hanushek@stanford.edu

Ludger Wößmann
Ifo Institute for Economic Research
at the University of Munich and CESifo
Poschingerstr. 5
81679 Munich, Germany
Phone: (+49) 89 / 9224-1699
woessmann@ifo.de

* This research was supported by CESifo under the project “International Educational Performance”.

1. Introduction

Many countries worry about the relative merits of a selective versus comprehensive school system, and the resulting system choices are surprisingly different. Some countries track students into differing-ability schools as early as at age 10 (e.g., Austria, Germany, Hungary, and the Slovak Republic). By contrast, others including Canada, Japan, Norway, Sweden, the United Kingdom, and the United States essentially keep their entire lower secondary school system comprehensive. Parents and politicians alike would like to know whether it has consequences for the equity and efficiency of educational outcomes if a country tracks its students into different school types, hierarchically structured by performance. Such macro issues of institutional structure are extraordinarily difficult to evaluate within individual countries, largely because the variations in structure that exist there are almost certainly related to the characteristics of the families and schools choosing to follow an anomalous pattern. To deal with these analytical complexities, we provide evidence from international experiences across countries.

The arguments about school placement policies – variously called tracking, streaming, or ability grouping – often rest on a perceived trade-off between equity and efficiency.[1] Some discussions of tracking are mainly concerned with placements between different types of schools and others with placements into different tracks within schools, but the arguments for and against tracking are basically the same.[2] The central argument behind tracking is that homogeneous classrooms permit a focused curriculum and appropriately paced instruction that leads to the maximum learning by all students. In such a situation, the teacher does not have to worry about boring the fastest learners or losing the slowest learners. The arguments for ungrouped classrooms largely revolve around concerns that the lower groups will be systematically disadvantaged by slower learning environments that leave them far behind the skills of those in the upper groups. The argument frequently goes further to relate preparation on entry into school to socio-economic background of the students, implying that grouping will also lead to continuing bias against more disadvantaged students.

---

[1] It appears that the costs of tracked and untracked systems are roughly comparable. Therefore, although we do not perform any direct efficiency calculations, we often refer to variations in outcomes in the loose manner of efficiency differences.

[2] See the papers on "comprehensive and selective schooling" collected in Heath (1984) for examples of the UK-based discussion of streaming between schools and Slavin (1990) for an example of the US-based discussion of ability grouping within schools.

The argument in favor of or against tracking gets even more complicated once possible peer effects are taken into account, because the precise nature of any interactions then becomes a key element in considering tracking. Proponents of ungrouped classrooms often suggest that heterogeneous classrooms might give rise to efficiency gains through nonlinear peer effects: the higher ability students lose nothing, but the lower ability students gain through the interaction (from motivation, better classroom discussion, and the like). By contrast, if the impact of peer achievement is linear, tracking would tend to increase the variance in outcomes without having any clear impact on the level of achievement (e.g., Argys et al. 1996). And if individuals are better off with peers of their own ability level, tracking could even improve the level of performance while possibly also reducing inequality (e.g., Dobbelsteen et al. 2002).[3] Thus, theory suggests considerable uncertainty about the impact of tracking on both the level and distribution of schooling outcomes.[4]

So far, the empirical literature attempting to sort out the effects of tracking on both the level and distribution of outcomes has followed two general strategies. The difficulty for any empirical research is that the major elements of the institutional structure of schools are choices whose impact is difficult to separate from other influences on achievement. When some schools or local education authorities introduce alternative structures, these choices are likely to be linked to other features of the students and schools if for no other reason than parental choices of residence and schools. Thus, the first empirical approach, which focuses on tracking within schools, attempts to standardize for heterogeneity across institutional structures through statistical analyses of measured factors (see Argys et al. 1996; Betts and Skolnick 2000; Betts et al. 2003; Figlio and Page 2002).[5]

Alternatively, if operating at the level of nations or states, the lack of within-state variation eliminates any control group unless there is variation over time. Thus, the second empirical approach, which focuses on tracking between different types of schools, looks within countries for situations where the institutional structure is altered and with some

---

[3] Lazear (2001) provides an alternative model of possible externalities within classrooms that lead to nonlinear effects of peer composition on student outcomes, which also generally implies efficiency improvements through grouping.

[4] For recent advanced theoretical treatments of the effects of tracking, see Brunello and Giannini (2004); Epple et al. (2002); and Meier (2004).

[5] The direct analyses of tracking are also supplemented by investigations of peer achievement effects. Early peer investigations were not very concerned about problems of omitted variables and simultaneity (i.e., the "reflection problem"). More recent peer studies have concentrated on those issues (Hanushek et al. 2003; Hoxby 2000). Nonetheless, the importance of peer ability remains disputed.

embellishments compares outcomes before and after.[6] The results of the different empirical analyses, while far from uniform, tend to suggest that tracking leads to more inequality in outcomes, particularly from the perspective of family backgrounds (but see Figlio and Page 2002 for an opposite finding).

The concern with both empirical approaches is that other unmeasured factors bias the estimated impacts of tracking. For example, with the trend analyses, the change in tracking structure is frequently just one of a series of changes to the schools. While these studies also include a variety of controls for other observable factors, it is hard to assess whether they sufficiently capture the concomitant factors that might affect student outcomes over time. The statistical analyses of tracking that employ both national and local samples for U.S. schools face complications of family residential choice plus generally sparse controls for family, teacher, and school differences – elements that are likely both to affect achievement and to be related to the institutional structure of classrooms.

To address these empirical problems, we use the macro variation in both the institutional structure of between-school tracking and student performance that exists across countries to sort out the impacts of tracking. Of course many other things also differ by country, leading us to adopt a differences-in-differences strategy to parse the effects of tracking. In this, we compare the level and distribution of performance of younger students (before tracking is introduced in any country) with those of older students (after some countries have started tracking) across countries with and without tracking, effectively using early outcomes in each country as the control. The existence of several large international assessment programs permits a consistent evaluation of student performance across a wide range of countries.

Our analysis provides reasonably strong support for the disequalizing effects of early tracking. Variation in performance, measured in a variety of ways, tends to increase across levels of schooling when a country employs early tracking. On the other hand, the evidence about possible efficiency gains from tracking is more mixed.

The remainder of the paper is structured as follows. Section 2 discusses the empirical identification strategy in detail. Section 3 describes the data. Section 4 presents the results on the impact of tracking on educational inequality, mean performance, and gainers and losers in the performance distribution. Section 5 concludes.

---

[6] For analyses of structural changes in the United Kingdom and Sweden, see Dearden et al. (2002); Galina-Rueda and Vignoles (2004); Harmon and Walker (2000); and Meghir and Palme (2004).

2. Cross-Country Identification

Understanding the impacts of macro institutional factors requires observing instances both of use of the structure and nonuse. In the case of between-school tracking, with the rare exception of when a country changes policies, the institution is common to all of the schools, implying that variation within countries is not useful.[7] At the same time, international comparisons face monumental problems because of the heterogeneity of nations. Quite obviously, finance and operations of school systems as well as social structure, family backgrounds, and a host of other, often unobserved factors besides tracking affect the observed outcomes.

Consider a simple model:

$$A_{ig}^{c} = \alpha_{c} + \gamma T_{ig}^{c} + X_{ig}^{c}\beta + \varepsilon_{ig}^{c} \quad (1)$$

where individual achievement of student *i* in grade *g* and country *c* (A_{ig}^{c}) is determined by a country specific intercept (α), varying attributes of families and schools (X), the existence of tracking (T), and an error (ε). In principle, if we could measure the various inputs to achievement, we could directly estimate equation (1). Two problems exist, however. First, we do not have sufficient knowledge or data to be confident of any estimates of the β (see Hanushek 2003). Second, with respect to the influences of tracking, if every student in the country is subject to tracking, T will be a constant, and we cannot estimate its influence on achievement.

In reality, no country tracks students between differing-ability schools in the early primary grades. Thus, we can consider looking at the changes that occur between primary school (grade *g*) and later schooling (grade g^{*}). A simple estimate of the impact of tracking could be found by looking at the average difference in achievement between *g* and g^{*} for a country that introduced tracking during that period:

$$\overline{\Delta A}^{c} = \gamma + \left(\overline{\Delta X}^{c}\beta + \overline{\Delta \varepsilon}^{c}\right) \quad (2)$$

In principle, if none of the X's changed much and if the change in average errors had an expected value of zero, we could estimate the impact of tracking (γ) simply by observing the growth in achievement over time for a single country.

---

[7] In the US, the use of magnet schools with specialized curriculum does vary across cities. These programs have not been evaluated very thoroughly, and both their existence and the selection rules for students is often closely related to their use as a device for the racial desegregation of schools.

It is nonetheless implausible to believe that all systematic influences across grades and across different tests are irrelevant to achievement. Specifically, the normal pattern of achievement gains between g and g^* would be intertwined with the impact of tracking. To deal with this, we can compare the growth in achievement across tracked countries and untracked countries, where the countries without tracking indicate the expected achievement gain in the absence of tracking:

$$\gamma = \overline{\overline{\Delta A}}_{tracked} - \overline{\overline{\Delta A}}_{untracked} + \left(\overline{\overline{v}}_{tracked} - \overline{\overline{v}}_{untracked}\right) \quad (3)$$

The impact of tracking can then be estimated by comparing the average achievement gain in tracked countries to that in untracked countries (where the double bar indicates averages across the groups of countries).

The estimation still depends upon the expected composite errors (v) being uncorrelated with the existence of tracking. This would be violated if, for example, the observed tests came from widely different cohorts of students such that the X's were to change (and to be correlated across countries with the existence of tracking), or if tracked nations tended to introduce more changes in their schools between the testing of students in different grades. We return to this below.

In reality, we estimate equation (3) in a regression framework where mean performance in grade g^* is regressed on mean performance in grade g along with an indicator for the existence of tracking. Thus, our approach applies a differences-in-differences methodology to the cross-country comparisons, combining tests in primary school with tests in secondary school. The effect of tracking is identified by comparing performance differences between primary and secondary school across tracked and non-tracked systems, where each country's own primary-school outcome is used as a control for its secondary-school outcome.

We also estimate a similar equation for inequality in performance. The simplest model is one where the variation in outcomes within countries are magnified (or shrunk) by the use of tracking. Again, the most basic model is a regression of late variance on early variance plus an indicator for tracking.

3. School Performance Data

International testing of students began in the early 1960s when the International Association for the Evaluation of Educational Achievement (IEA) developed a mathematics

test that could be used to compare student performance across countries. Although the earliest testing was plagued by uncertainties about the within-country sampling, the selectivity of students who were not in school, and a variety of other factors, more recent testing has followed strict protocols with elaborate efforts to ensure both high quality test designs and representative sampling of students.

To implement the differences-in-differences estimation, we concentrate on the series of international assessments conducted since 1995. We match international student achievement tests in secondary school with tests late in primary school. Because the methodology requires a stable educational system, we concentrate on roughly contemporaneous measures of performance at the two different grade levels.[8] We supplement the six different test observations that meet this requirement, however, by following the 1995 cohort of 4th grade students that subsequently was assessed in the 8th grade in 1999 (on the TIMSS math and science tests). Table 1 summarizes the comparisons that are used, and the data and sources are described in detail in the Appendix.

Tests are found in reading, mathematics, and science. Each assessment produces 18 to 26 country level observations. For analytical purposes, the differences in the tests and subjects lead us to treat each of the eight assessment pairs as a separate test of the impacts of early tracking, although the common grouping of countries implies that these are not truly independent tests.

In our analyses, we use the data on age of first tracking as a dummy representing whether an education systems tracks its students before the age at which the specific secondary-school test is performed or not. For the PISA secondary-school tests, we consider tracking by age 15 (the average student age on the two PISA tests is 15 years and 9 months); for the TIMSS secondary-school tests, we consider tracking by age 14 at the latest (corresponding to an average testing age of 14 years and 5 months). Half the countries in our samples based on the PISA tests had a tracked system by the age of 15. The share of countries that tracked by the age of 14 in the TIMSS tests is roughly one third (see Appendix Table A2), reflecting both the earlier testing age and the different country compositions of the samples.

[8] Comparing different cohorts at one point in time minimizes any contamination of variations in other school policies, but it does so at the cost of any inherent variation in family background and peers that exists across different cohorts. Although we also follow a single cohort (see below), we emphasize comparisons at a given time because we believe that school policies tend to be more volatile than family backgrounds of cohorts.

Table 1: Matching Pairs of International Tests in Primary and Secondary School

	Secondary-school test			Primary-school test			Subject	Joint
	Test	Year	Grade/Age	Test	Year	Grade/Age		countries
1.	PISA	2003	15-year-olds	PIRLS	2001	4th grade	Reading	18
2.	PISA	2000/02	15-year-olds	PIRLS	2001	4th grade	Reading	20
3.	TIMSS	1995	8th grade	TIMSS	1995	4th grade	Math	26
4.	TIMSS	1995	8th grade	TIMSS	1995	4th grade	Science	26
5.	TIMSS	2003	8th grade	TIMSS	2003	4th grade	Math	25
6.	TIMSS	2003	8th grade	TIMSS	2003	4th grade	Science	25
7.	TIMSS	1999	8th grade	TIMSS	1995	4th grade	Math	18
8.	TIMSS	1999	8th grade	TIMSS	1995	4th grade	Science	18

Notes: PISA = Program for International Student Assessment. – PIRLS = Progress in International Reading Literacy Study. – TIMSS = Trends in International Mathematics and Science Study (formerly Third International Mathematics and Science Study).

4. Impacts of Early Tracking

Because of the importance attached to inequality in the existing literature, we begin with an analysis of distributional aspects of tracking. This is followed by implications for mean performance.

4.1 Tracking and Inequality

The nature of the international comparisons and the relationship with tracking is easiest to see in the data on inequality for the most recent comparison: reading performance on the 2003 administration of the PISA test for 15 year olds compared to the 2001 administration of the PIRLS test for 4th graders. Figure 1 plots the relative standard deviation of scores for countries with early tracking (solid lines) versus countries without early tracking (dashed lines).[9]

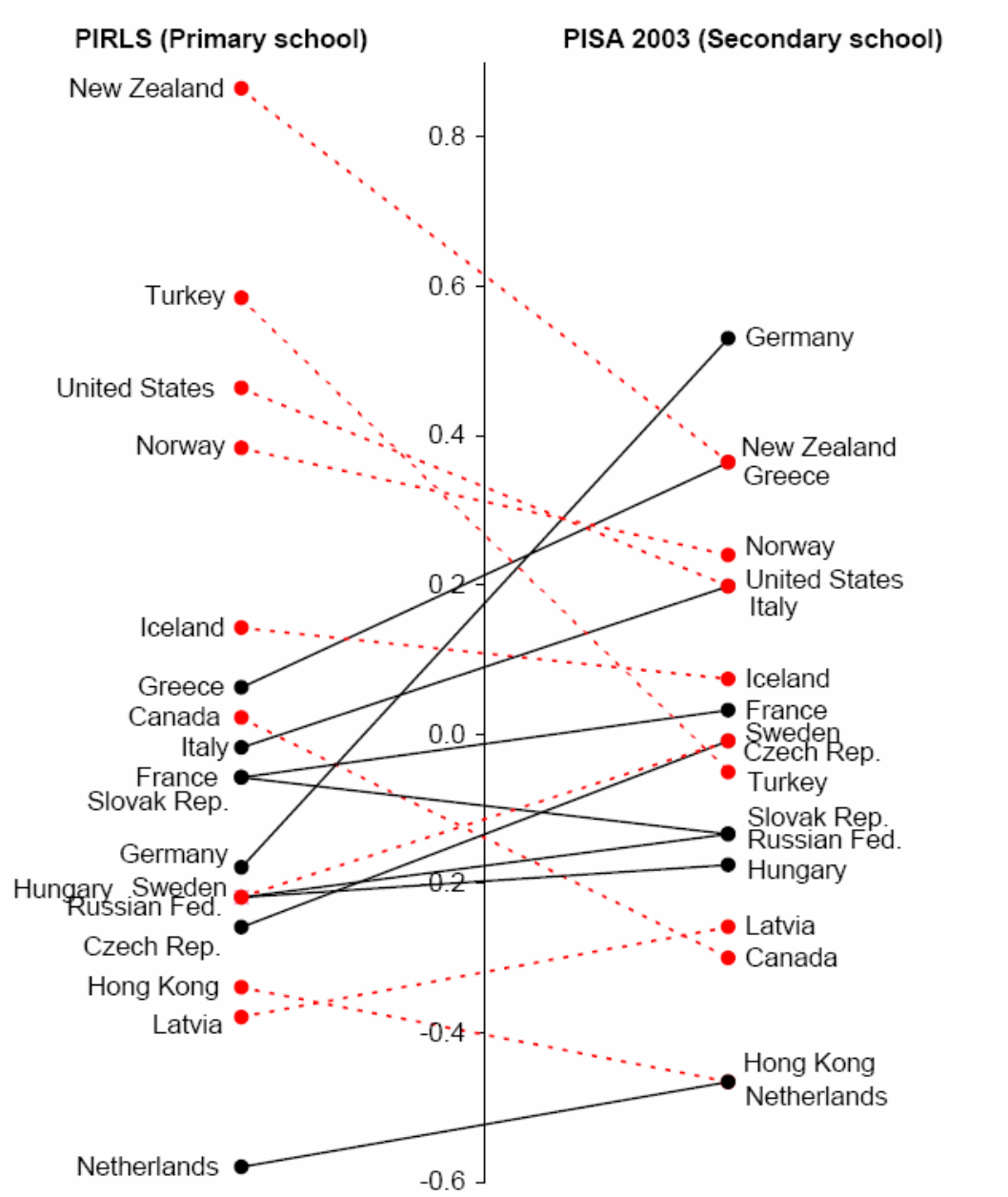
Relative inequality increases in every country with tracking except the Slovak Republic, while relative inequality decreases in every country without tracking except for Sweden and Latvia. Out of the 18 countries, the top four countries in terms of the increase in inequality between primary and secondary school are all early trackers (Germany with an increase of 0.71, Greece 0.30, Czech Republic 0.25, and Italy 0.22). The bottom six countries with the largest decrease in inequality are all late trackers that do not track before the age of PISA testing (Turkey -0.63, New Zealand -0.50, Canada -0.32, United States -0.27, Norway -0.14, and Hong Kong -0.13).

The regression analysis expands this to consider different measures of inequality: the standard deviation of test scores within each country; the test-score difference between the student performing at the 75th percentile and the student performing at the 25th percentile in each country; and the performance difference between the 95th and the 5th percentile. We also provide a comparison with estimation of a simple model of average achievement that, along the lines of equation (1), compares mean performance of the 15 year olds just to tracking.

As the results reported in columns (1), (3), and (5) of Table 2 show, none of the three inequality measures is statistically significantly related to tracking in a simple bivariate analysis. However, as argued in Section 2, these bivariate estimates may be biased by general heterogeneity in inequality of the participating countries. Thus, columns (2), (4), and (6) report differences-in-differences estimates of the effect of early tracking on the three

[9] Standard deviations are expressed relative to the average national standard deviation on each test.

Figure 1: Inequality in Primary and Secondary School



Notes: Standard deviation of test scores in the national population (difference to international average of national standard deviations in each test). – Countries with a tracked school system before the age of 16 have solid lines, countries without tracking before age 16 have dashed lines.

Table 2: Tracking and Inequality: PISA 2003 and PIRLS

	(1)	(2)	(3)	(4)	(5)	(6)
Measure of inequality:	Standard deviation		75^{th}-25^{th} percentile difference		95^{th}-5^{th} percentile difference	
Early tracking	0.046	0.248**	0.157	0.385*	0.143	0.834**
	(0.136)	*(0.110)*	*(0.206)*	*(0.185)*	*(0.462)*	*(0.376)*
Inequality in primary school (measure: see top row)		0.594***		0.538**		0.605***
		(0.129)		*(0.197)*		*(0.127)*
Constant	3.970***	2.165***	5.298***	3.156***	13.027***	6.957***
	(0.092)	*(0.399)*	*(0.153)*	*(0.781)*	*(0.316)*	*(1.332)*
Number of countries	18	18	18	18	18	18
R^2	0.007	0.479	0.035	0.366	0.006	0.506

Dependent variable: Inequality in secondary school, as indicated in top row. – Huber-White heteroscedasticity-consistent standard errors in parentheses. – Significance levels: *** 1 percent. – ** 5 percent. – * 10 percent.

inequality measures which condition on the extent of educational inequality already present in late primary school, before tracking in any country. With all three measures of inequality, it is obvious that countries that exert high inequality already in primary school also tend to have high inequality in secondary school. The point estimates of roughly 0.6 indicate that schools everywhere tend to reduce the inequality which was present in primary grades – and which presumably represents the proportionately greater influence of families.

More importantly, on all three measures of inequality, countries that track their students before age 15 show a statistically significantly larger inequality on the PISA 2003 secondary-school test, once the difference in inequality that existed already in primary school is accounted for. Specifically, early trackers show a national standard deviation of test scores in secondary school that is one quarter of a cross-country standard deviation larger than non-trackers. Consider for example the observed country differences in outcome variation. The minimum national standard deviation of 3.5 (Hong Kong and the Netherlands) is noticeably different from the maximum national standard deviation of 4.5 (Germany) on the PISA 2003 test. The results suggest that the effect of early tracking can account for one quarter of the difference in inequality between the most inequitable and the most equitable country.

Figure 1 makes apparent why simple bivariate estimates do not reveal this pattern: None of the five countries with the largest inequality in primary school (New Zealand, Turkey, United States, Norway, and Iceland) have early tracking of students. Across the countries, the correlation between the national standard deviation in primary school and the early-tracking dummy is -0.472 (statistically significant at the 5 percent level).

The estimates across the other seven pairs of international achievement tests are generally consistent with the results in Table 2 but are not as strong or statistically significant. Table 3 reports the differences-in-differences results using the standard deviation as the inequality measure.[10] With the exception of the PISA 2000/02-PIRLS pair, all estimates of the coefficient on early tracking are positive, and four are statistically significant at the 10 percent level or better. For the insignificant results of columns (11)-(13), inequality in secondary school is not even statistically significantly related to inequality in primary school, raising some concerns about the specific tests.

The limited samples of countries preclude very elaborate specification checks, but some extensions are interesting. First, rather than entering the tracking variable as a dummy,

[10] Given the consistency across measures of inequality, we report only the results for standard deviations. The results for the other two measures were qualitatively very similar.

Table 3: Tracking and Inequality: Different Tests

	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Secondary-school test:	PISA 03	PISA 00/02	TIMSS 95	TIMSS 95	TIMSS 03	TIMSS 03	TIMSS 99	TIMSS 99
Primary-school test:	PIRLS	PIRLS	TIMSS 95	TIMSS 95	TIMSS 03	TIMSS 03	TIMSS 95	TIMSS 95
Subject:	Reading	Reading	Math	Science	Math	Science	Math	Science
Early tracking	0.248**	-0.018	0.147*	0.197**	0.013	0.105	0.005	0.208*
	(0.110)	*(0.077)*	*(0.076)*	*(0.084)*	*(0.054)*	*(0.073)*	*(0.074)*	*(0.107)*
Inequality in primary school (standard deviation)	0.594***	0.255*	0.476	0.843***	-0.014	0.252	0.099	0.785***
	(0.129)	*(0.139)*	*(0.306)*	*(0.224)*	*(0.248)*	*(0.176)*	*(0.146)*	*(0.135)*
Constant	2.165***	1.829***	0.811	0.675	1.323***	1.073***	1.532***	0.869***
	(0.399)	*(0.325)*	*(0.499)*	*(0.411)*	*(0.291)*	*(0.189)*	*(0.255)*	*(0.275)*
Number of countries	18	20	26	26	25	25	18	18
R^2	0.479	0.255	0.258	0.419	0.003	0.257	0.022	0.558

Dependent variable: Inequality in secondary school, measured by the standard deviation in test scores. – Huber-White heteroscedasticity-consistent standard errors in parentheses. – Significance levels: *** 1 percent. – ** 5 percent. – * 10 percent.

we can also enter tracking as a linear variable depicting the age at which a country first tracks its students. Unfortunately, the continuous variation in when the tracking occurs is limited, with no country starting to track at the age of 13, for example. Results using the linear tracking variable (available from the authors) are broadly consistent with results using the simple existence of tracking, and the main impact comes from the mere existence of early tracking with no consistent linear pattern detectable for the age at which tracking occurred.

Additionally, experimentation with adding further control variables to the estimation did not change the basic results. In terms of the estimates of Table 2, adding GDP per capita and/or a country's cumulative educational expenditure per student by age 15 left the impact of tracking largely unchanged.[11]

4.2 Tracking and Mean Performance

Given that comprehensive schooling systems seem to reduce inequality, the question arises whether this effect is achieved by improving the lowest performers or by holding back the best performers. That is, does performance converge at a lower or higher level? We first estimate the effect of tracking on a country's mean performance level using the same differences-in-differences identification strategy as before; following that, we estimate the effects at different percentiles of student performance in the next section.

Table 4 reports the results on the effect of early tracking on mean performance for all 8 pairs of international student achievement tests. In all pairs, we see a clear tendency for countries which performed better on average in primary school to also perform better in secondary school.

The impact of early tracking is, however, inconsistent across subjects and tests. The two reading comparisons indicate a statistically significant lower achievement associated with early tracking. Similarly, the mathematics results are always lower with early tracking, although the result is statistically significant at the 10 percent level or better in only one of the three comparisons. For science, however, two of the three estimates indicate positive achievement effects from early tracking (and one is statistically significant at the 5 percent level).

[11] OECD (2004) reports GDP per capita (in purchasing power parities) for 15 of the 18 countries. When included, it enters statistically significantly positive, while the tracking dummy also remains statistically significantly positive. The expenditure measure (again in purchasing power parities) is available for 13 countries but does not enter significantly, although the significance level of the tracking dummy falls to 15 percent.

Table 4: Tracking and Mean Performance

	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
Secondary-school test:	PISA 03	PISA 00/02	TIMSS 95	TIMSS 95	TIMSS 03	TIMSS 03	TIMSS 99	TIMSS 99
Primary-school test:	PIRLS	PIRLS	TIMSS 95	TIMSS 95	TIMSS 03	TIMSS 03	TIMSS 95	TIMSS 95
Subject:	Reading	Reading	Math	Science	Math	Science	Math	Science
Early tracking	-1.053***	-0.951***	-0.062	0.597**	-0.021	-0.013	-0.410*	0.234
	(0.343)	*(0.287)*	*(0.135)*	*(0.222)*	*(0.157)*	*(0.161)*	*(0.219)*	*(0.370)*
Mean performance in primary school	0.676***	0.643***	0.965***	0.738***	0.928***	0.929***	1.045***	0.828***
	(0.139)	*(0.130)*	*(0.063)*	*(0.097)*	*(0.085)*	*(0.075)*	*(0.088)*	*(0.124)*
Constant	0.526**	0.475**	0.019	-0.184*	0.006	0.004	0.137	-0.078
	(0.230)	*(0.203)*	*(0.087)*	*(0.103)*	*(0.101)*	*(0.095)*	*(0.079)*	*(0.102)*
Number of countries	18	20	26	26	25	25	18	18
R^2	0.582	0.635	0.900	0.779	0.858	0.863	0.921	0.751

Dependent variable: Mean performance in secondary school. – Huber-White heteroscedasticity-consistent standard errors in parentheses. – Significance levels: *** 1 percent. – ** 5 percent. – * 10 percent.

As an alternative approach, we allow for the possible correlation of the residuals of the inequality and the mean-performance equations. In order to improve the estimation efficiency, we estimate the two equations by seemingly unrelated regressions (SUR). The results in Appendix Table A3 reveal no change in any of the substantive previous results, only lifting the significance level of the effect of early tracking on inequality in the TIMSS 2003 science test.

4.3 Who Gains, Who Loses?

One final issue is where any losses (or gains) from early tracking are found in the distribution. To address this, we estimate the effect of early tracking on the performance of students at different percentiles of the performance distribution, again in differences-in-differences models. Specifically, we estimate whether a student at the 5^{th} percentile (or 25^{th}, 75^{th}, and 95^{th} percentiles) of the national distribution is affected by tracking. Although effects cannot be statistically significantly estimated in most pairs of international achievement tests, where they can, they reinforce the results in Tables 2-4.

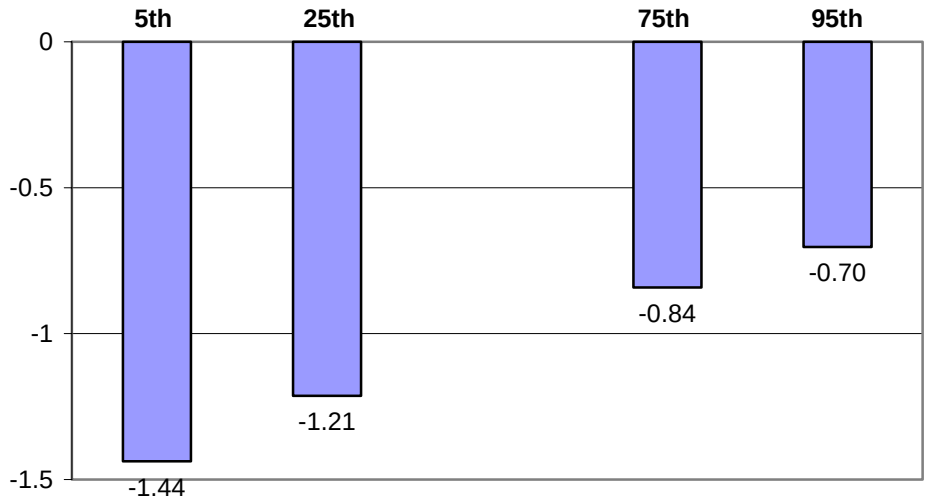
For example, the increased inequality and decreased mean performance in tracked systems detected in the PISA 2003-PIRLS pair come from the lower percentiles losing more than the upper ones, even though each of the four percentiles loses a statistically significant amount. The coefficient estimates on the early-tracking dummy for the different achievement levels are depicted in Figure 2, which shows that lower performers suffer more from early tracking than higher ones.

Across the estimates from the remaining samples (available from the authors), the most striking finding is that in no case do some students gain at the expense of others; both high and low achievers lose (or, in the one case of a positive effect on mean performance, gain) from tracking. The net impact comes from the differential impacts on different parts of the distribution.

5. Conclusion

This analysis provides preliminary results about the impact of early tracking on the level and distribution student performance. The results consistently indicate that early tracking increases inequality in achievement. Although the evidence on the level of

Figure 2: The Effect of Tracking on Performance at Different Percentiles: PISA 2003 and PIRLS



Note: Coefficient estimate on the early-tracking dummy in separate differences-in-differences estimations of the performance of the X^{th} percentile in PISA 2003 on the performance of the X^{th} percentile in PIRLS and the early-tracking dummy.

performance is less certain, there is very little evidence that there are efficiency gains associated with this increased inequality.

On the research side, these preliminary results also suggest the value of further study of tracking. Some of the literature has suggested that one channel for increasing inequality is reinforcing the effects of family background. Specifically, if much of the early inequality in achievement is associated with differences in family background, many of the track placements will be associated directly with family background. Indeed, some have suggested that family background is a driving force in setting track placements even beyond its impact on early achievement levels (e.g., Schnepf 2003). The implications for family background inequality can potentially be investigated through use of the micro data generated by the international assessments. Beyond that, with the micro data it would be possible to consider more fully the underlying structural model of achievement that would generate these patterns of aggregate outcomes. Also, extending the dichotomous analysis between tracked and non-tracked systems pursued in this paper, there may be heterogeneity in the rigidity of tracked systems. Future research may explore the extent to which allowing mobility across tracks might reduce the negative effects of tracking.

From a policy perspective, it seems incumbent on those advocating early tracking in schools to identify the potential gains from this. These preliminary results suggest that countries lose in terms of the distribution of outcomes, and possibly also in levels of outcomes, by pursuing such policies.

Appendix: Data Sources and Description

The most recent international test employed is the 2003 edition (data release: December 2004) of the Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development (OECD).[12] PISA tested representative samples of 15-year-old students in reading, math, and science, with a focus of test items on real-life applications. A recent primary-school test to which the PISA test can be matched is the Progress in International Reading Literacy Study (PIRLS). In 2001, the International Association for the Evaluation of Educational Achievement (IEA) conducted the PIRLS reading test to 4th-grade students,[13] which is the grade just before the first countries start tracking their schools. There are 18 countries that participated both in PISA 2003 and in PIRLS. Appendix Table A1 provides a list of countries participating in each pair of tests.

Since the mid-1990s, there are seven further international student achievement tests at the end of lower secondary education to which we can match specific primary-school tests, all of which tested representative samples of students in each participating countries (see Table 1). The first PISA study, also testing 15-year-olds in reading, math, and science, was conducted in 2000 for most participating countries and in 2002 for several additional countries. We match the PISA 2000/02 test again with the 2001 PIRLS primary-school test, which gives a sample of 20 countries participating in both tests. Next, the IEA performed the Third International Mathematics and Science Study (TIMSS, later re-named to Trends in International Mathematics and Science Study) in 1995, which tested both 4th-grade and 8th-grade students in math and science.[14] Matching the TIMSS 1995 tests in primary and secondary school, there are 26 countries participating both in the two math tests and in the two science tests. The next primary-school TIMSS tests were conducted in 2003, which we can match to the TIMSS 2003 secondary-school tests, yielding a sample of 25 countries participating in primary and secondary school both in math and in science.

All these matches test primary- and secondary-school students at exactly or roughly the same point in time. We can also follow specific cohorts of students over time. This is possible by relating the 8th-grade performance on the TIMSS tests in 1999 to the 4th-grade performance on the TIMSS tests in 1995. That is, the very same cohort which was tested in math and science in 4th grade in 1995 was again tested in 8th grade in 1999. 18 countries participated both in the 1995 4th-grade and in the 1999 8th-grade math and science tests, allowing for matching of representative samples from the same cohort followed over time.

We take the data on means, standard deviations, and percentiles of the test-score performance on the different international tests from the following sources: OECD (2003; 2004) for reading performance in PISA 2000/2002 and PISA 2003; Mullis et al. (2003) for reading performance in PIRLS; Beaton et al. (1996a; 1996b) for secondary-school math and science performance in TIMSS 1995; Mullis et al. (1997; 2000; 2004) for math performance in TIMSS 1995 primary school, TIMSS 1999 and TIMSS 2003; and Martin et al. (1997; 2000; 2004) for science performance in TIMSS 1995 primary school, TIMSS 1999 and TIMSS 2003.

For the purposes of this paper, we re-scale the test scores of each primary-secondary pair of tests so that they are normalized to have a mean of zero and a cross-country standard deviation of one between the countries jointly participating at both test levels. Note that this normalization refers to the cross-country variation of test scores among the sample of participating countries only, without considering possible differences in the within-country variation of test scores between the primary- and secondary-school tests. Thus, for example, the mean of the standard deviation of test scores *within each country* is considerably larger in the PISA 2003 test than in the PIRLS test, at 4.0 versus 2.9 *cross-country* standard variations on each of the tests, respectively. That is, in PISA 2003, the standard deviation of test scores *within* a country was, on average, four times as large as the standard deviation of test scores *across* the 18 countries.

We collected data on the age at which students are tracked into different schools for the first time in each country from different sources, including the data collections of the European Commission (2000; 2002), the Encyclopedia of national education systems of Postlethwaite (1996), a table in OECD (2003), and detailed

---

[12] TIMSS assessments for 2003 were also released in December 2004, but we start with the PISA 2003 test because it tests students who are older and thus longer exposed to tracking than the students tested in the TIMSS tests and because it has a broader coverage of developed countries.

[13] Specifically, PIRLS tested the upper of the two adjacent grades with the largest share of nine-year-olds in each country, which is usually fourth grade.

[14] Specifically, the different TIMSS tests tested the upper of the two adjacent grades with the largest share of 9-year-olds (4th grade) and 13-year-olds (8th grade), respectively, in each country. The first TIMSS test also tested the lower of each of these grades (3rd and 7th grade), but we stick with 4th and 8th grade because 4th grade is just before the first countries start tracking their schools and 8th grade allows more time for tracking to exert its effects.

country-specific inquiries. The mean age of students at the time of first tracking across the 45 countries considered in this paper is 15.2, ranging from a minimum of 10 to a maximum of 18.

Descriptive statistics for the eight pairs of international tests are presented in Appendix Table A2.

Appendix References

Beaton, Albert E., Ina V.S. Mullis, Michael O. Martin, Eugenio J. Gonzalez, Dana L. Kelly, Teresa A. Smith (1996a). *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Beaton, Albert E., Michael O. Martin, Ina V.S. Mullis, Eugenio J. Gonzalez, Teresa A. Smith, Dana L. Kelly (1996b). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

European Commission (2000, 2002). *Key Data on Education in Europe*. Luxembourg: Office for Official Publications of the European Communities.

Martin, Michael O., Ina V.S. Mullis, Albert E. Beaton, Eugenio J. Gonzalez, Teresa A. Smith, Dana L. Kelly (1997). *Science Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Martin, Michael O., Ina V.S. Mullis, Eugenio J. Gonzalez, Kelvin D. Gregory, Teresa A. Smith, Steven J. Chrostowski, Robert A. Garden, Kathleen M. O'Connor (2000). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: Boston College.

Martin, Michael O., Ina V.S. Mullis, Eugenio J. Gonzalez, Steven J. Chrostowski (2004). *TIMSS 2003 International Science Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grade*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, Ina V.S., Michael O. Martin, Albert E. Beaton, Eugenio J. Gonzalez, Dana L. Kelly, Teresa A. Smith (1997). *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Mullis, Ina V.S., Michael O. Martin, Eugenio J. Gonzalez, Kelvin D. Gregory, Robert A. Garden, Kathleen M. O'Connor, Steven J. Chrostowski, Teresa A. Smith (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: Boston College.

Mullis, Ina V.S., Michael O. Martin, Eugenio J. Gonzalez, Ann M. Kennedy (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary School in 35 Countries*. Chestnut Hill, MA: International Study Center, Boston College.

Mullis, Ina V.S., Michael O. Martin, Eugenio J. Gonzalez, Steven J. Chrostowski (2004). *TIMSS 2003 International Mathematics Report: Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grade*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Organisation for Economic Co-operation and Development (OECD) (2003). *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) (2004). *Learning for Tomorrow's World: First Results from PISA 2003*. Paris: OECD.

Postlethwaite, T. Neville, ed. (1996). *International Encyclopedia of National Education Systems*. 2nd Edition. Oxford: Pergamon.

Table A1: Countries Participating in Each Pair of Tests

PISA 2003 + PIRLS	PISA 2000/02 + PIRLS	TIMSS 1995 8th grade + TIMSS 1995 4th grade	TIMSS 2003 8th grade + TIMSS 2003 4th grade	TIMSS 1999 8th grade + TIMSS 1995 4th grade
Canada	Argentina	Australia	Armenia	Australia
Czech Republic	Bulgaria	Austria	Australia	Canada
France	Canada	Canada	Belgium (Flemish)	Cyprus
Germany	Czech Republic	Cyprus	Cyprus	Czech Republic
Greece	France	Czech Republic	England	England
Hong Kong	Germany	England	Hong Kong	Hong Kong
Hungary	Greece	Greece	Hungary	Hungary
Iceland	Hong Kong	Hong Kong	Iran	Iran
Italy	Hungary	Hungary	Italy	Israel
Latvia	Iceland	Iceland	Japan	Japan
Netherlands	Israel	Iran	Latvia	Korea
New Zealand	Italy	Ireland	Lithuania	Latvia
Norway	Latvia	Israel	Moldova	Netherlands
Russian Federation	Macedonia	Japan	Morocco	New Zealand
Slovak Republic	New Zealand	Korea	Netherlands	Singapore
Sweden	Norway	Kuwait	New Zealand	Slovenia
Turkey	Russian Federation	Latvia	Norway	Thailand
United States	Sweden	Netherlands	Philippines	United States
	United Kingdom	New Zealand	Russian Federation	
	United States	Norway	Scotland	
		Portugal	Singapore	
		Scotland	Slovenia	
		Singapore	Taipei	
		Slovenia	Tunisia	
		Thailand	United States	
		United States		

Table A2: Descriptive Statistics for Each Pair of Tests

	National mean				National standard deviation		
	Mean	SD	Min	Max	Mean	Min	Max
Age of first tracking (45 countries)	15.222	2.679	10	18			
Pair 1. (18 countries)							
PISA 2003 reading	0	1	-2.040	1.570	3.993	3.528	4.524
PIRLS 2001 reading	0	1	-3.202	1.293	2.867	2.288	3.732
Early tracking	0.500						
Pair 2. (20 countries)							
PISA 2000/02 reading	0	1	-2.657	1.235	2.380	2.031	2.684
PIRLS 2001 reading	0	1	-2.952	1.092	2.200	1.778	2.954
Early tracking	0.500						
Pair 3. (26 countries)							
TIMSS 1995 8th-grade math	0	1	-2.350	2.258	1.604	1.065	2.001
TIMSS 1995 4th-grade math	0	1	-2.457	1.836	1.569	1.279	1.985
Early tracking	0.308						
Pair 4. (26 countries)							
TIMSS 1995 8th-grade science	0	1	-2.435	1.998	2.255	1.803	2.655
TIMSS 1995 4th-grade science	0	1	-2.678	1.575	1.802	1.432	2.105
Early tracking	0.308						
Pair 5. (25 countries)							
TIMSS 2003 8th-grade math	0	1	-1.997	1.863	1.311	1.020	1.700
TIMSS 2003 4th-grade math	0	1	-2.223	1.399	1.131	0.781	1.563
Early tracking	0.280						
Pair 6. (25 countries)							
TIMSS 2003 8th-grade science	0	1	-2.327	1.432	1.391	1.122	1.907
TIMSS 2003 4th-grade science	0	1	-2.515	1.043	1.143	0.723	1.977
Early tracking	0. 280						
Pair 7. (18 countries)							
TIMSS 1999 8th-grade math	0	1	-2.043	1.762	1.705	1.526	2.007
TIMSS 1995 4th-grade math	0	1	-2.384	1.711	1.743	1.441	2.172
Early tracking	0.333						
Pair 8. (18 countries)							
TIMSS 1999 8th-grade science	0	1	-2.075	1.375	2.433	2.013	3.019
TIMSS 1995 4th-grade science	0	1	-2.698	1.500	1.904	1.530	2.249
Early tracking	0.333						

Table A3: Tracking, Inequality and Mean Performance: Seemingly Unrelated Regressions

	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Secondary-school test:	PISA 03	PISA 00/02	TIMSS 95	TIMSS 95	TIMSS 03	TIMSS 03	TIMSS 99	TIMSS 99
Primary-school test:	PIRLS	PIRLS	TIMSS 95	TIMSS 95	TIMSS 03	TIMSS 03	TIMSS 95	TIMSS 95
Subject:	Reading	Reading	Math	Science	Math	Science	Math	Science
Dependent variable: Standard deviation in secondary school								
Early tracking	0.249**	-0.018	0.146*	0.197**	0.013	0.106*	0.004	0.212**
	(0.106)	*(0.068)*	*(0.080)*	*(0.085)*	*(0.056)*	*(0.063)*	*(0.060)*	*(0.086)*
Standard deviation in primary school	0.599***	0.248**	0.481**	0.847***	-0.021	0.266***	0.121	0.828***
	(0.147)	*(0.098)*	*(0.227)*	*(0.207)*	*(0.151)*	*(0.098)*	*(0.141)*	*(0.140)*
Constant	2.152***	1.844***	0.805**	0.668*	1.331***	1.058***	1.493***	0.785***
	(0.451)	*(0.224)*	*(0.356)*	*(0.380)*	*(0.175)*	*(0.118)*	*(0.246)*	*(0.275)*
R^2	0.479	0.255	0.258	0.419	0.003	0.257	0.021	0.557
Dependent variable: Mean performance in secondary school								
Early tracking	-1.057***	-0.952***	-0.042	0.554***	-0.023	-0.021	-0.323**	0.288
	(0.304)	*(0.264)*	*(0.157)*	*(0.211)*	*(0.167)*	*(0.162)*	*(0.155)*	*(0.249)*
Mean performance in primary school	0.685***	0.655***	0.947***	0.792***	0.935***	0.971***	0.961***	0.736***
	(0.156)	*(0.134)*	*(0.074)*	*(0.098)*	*(0.077)*	*(0.074)*	*(0.068)*	*(0.098)*
Constant	0.528**	0.476**	0.013	-0.170	0.007	0.006	0.108	-0.096
	(0.212)	*(0.186)*	*(0.078)*	*(0.111)*	*(0.087)*	*(0.086)*	*(0.083)*	*(0.141)*
R^2	0.582	0.635	0.900	0.776	0.858	0.861	0.915	0.743
Number of countries	18	20	26	26	25	25	18	18

The two equations in each column are jointly estimated with seemingly unrelated regressions (SUR). – Standard errors in parentheses. – Significance levels: *** 1 percent. – ** 5 percent. – * 10 percent.

References

Argys, Laura M., Daniel I. Rees, Dominic J. Brewer (1996). Detracking America's Schools: Equity at Zero Cost? *Journal of Policy Analysis and Management* 15 (4):623-645.

Betts, Julian R., Jamie L. Shkolnik (2000). The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools. *Economics of Education Review* 19 (1): 1-15.

Betts, Julian R., Andrew C. Zau, Lorien A. Rice (2003). *Determinants of Student Achievement: New Evidence from San Diego*. San Francisco: Public Policy Institute of California.

Brunello, Giorgio, Massimo Giannini (2004). Stratified or Comprehensive? The Economic Efficiency of School Design. *Scottish Journal of Political Economy* 51 (2): 173-193.

Dearden, Lorraine, Javier Ferri, Costas Meghir (2002). The Effect of School Quality on Educational Attainment and Wages. *Review of Economics and Statistics* 84 (1): 1-20.

Dobbelsteen, Simone, Jesse Levin, Hessel Oosterbeek (2002). The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition. *Oxford Bulletin of Economics and Statistics* 64 (1): 17-38.

Epple, Dennis, Elizabeth Newlon, Richard Romano (2002). Ability Tracking, School Competition, and the Distribution of Educational Benefits. *Journal of Public Economics* 83 (1): 1-48.

Figlio, David N., Marianne E. Page (2002). School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality? *Journal of Urban Economics* 51 (3): 497-514.

Galindo-Rueda, Fernando, Anna Vignoles (2004). The Heterogeneous Effect of Selection in Secondary Schools: Understanding the Changing Role of Ability. Paper presented at the CESifo/PEPG conference on "Schooling and Human Capital Formation in the Global Economy: Revisiting the Equity-Efficiency Quandary" in Munich, September.

Hanushek, Eric A. (2003). The Failure of Input-Based Schooling Policies. *Economic Journal* 113 (485): F64-F98.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, Steven G. Rivkin (2003). Does Peer Ability Affect Student Achievement? *Journal of Applied Econometrics* 18 (5): 527-544.

Harmon, Colm, Ian Walker (2000). The Returns to the Quantity and Quality of Education: Evidence for Men in England and Wales. *Economica* 67 (265): 19-35.

Heath, Anthony, ed. (1984). Comprehensive and Selective Schooling. Special Issue of the *Oxford Review of Education* 10 (1): 6-123.

Hoxby, Caroline M. (2000). Peer Effects in the Classroom: Learning from Gender and Race Variation. NBER Working Paper 7867. Cambridge, MA: National Bureau of Economic Research.

Lazear, Edward P. (2001). Educational Production. *Quarterly Journal of Economics* 116 (3): 777-803.

Meghir, Costas, Mårten Palme (2004). Educational Reform, Ability and Family Background. *American Economic Review*: forthcoming. (IFS Working Paper 04/10, London: Institute for Fiscal Studies.)

Meier, Volker (2004). Choosing between School Systems: The Risk of Failure. *Finanzarchiv* 60 (1): 83-93.

Schnepf, Sylke V. (2003). Inequalities in Secondary School Attendance in Germany. S^3RI Applications Working Paper A03/16. Southampton: Southampton Statistical Sciences Research Institute.

Slavin, Robert E. (1990). Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis. *Review of Educational Research* 60 (3): 471-499.