Middle School Mediocrity

Quality and Inequality in Secondary Education

Kristian Koerselman

koerselman@economistatwork.com
http://economistatwork.com

Cover
Kristian Koerselman

Publisher
Åbo Akademi University Press
Piispankatu 13, FI-20500 Turku, Finland
phone +358 (0)20 786 1468
forlaget@abo.fi
http://www.abo.fi/stiftelsen/forlag

Distributor
Oy Tibo-Trading Ab
PO Box 33, FI-21601 Pargas, Finland
phone +358 (0)2 454 9200
fax +358 (0)2 454 9220
tibo@tibo.net
http://www.tibo.net

Middle School Mediocrity

Middle School Mediocrity

Quality and Inequality in Secondary Education

Kristian Koerselman

Åbo 2011

Åbo Akademis Förlag – Åbo Akademi University Press

CIP Cataloguing in Publication

Koerselman, Kristian.
Middle school mediocrity : quality and inequality in secondary education / Kristian Koerselman – Åbo : Åbo Akademi University Press, 2011.
Diss.: Åbo Akademi University.
ISBN 978-951-765-620-7

ISBN 978-951-765-620-7
ISBN 978-951-765-621-4 (digital)
Painosalama Oy
Åbo 2011

Trust science, question the scientist.

Preface

Equality is one of those things which no good man could ever oppose. Call it mediocrity and the magic is lost.

Equality mostly exists in its negative form: inequality. We speak of a reduction of inequality rather than of an expansion of equality, suggesting that equality is something in which our society is necessarily lacking.

One can reduce inequality by transferring something from those who have more to those who have less. Sometimes, this can be done without a cost in terms of averages, but that does not mean that it is a Pareto improvement.

Common arguments for income redistribution do not necessarily hold when considering education. Furthermore, the lack of a good cardinal measure of educational achievement makes it harder to define efficiency. Inequality in educational outcomes is therefore not unambiguously bad, particularly when framed as the choice between a drive for excellence and a drive for mediocrity.

This thesis is centered around tracking, an educational policy which probably increases differences between children while keeping the median relatively unaffected. Tracking is hard to defend when framed as detrimental to equality, but it is also a guard against mediocrity – an effect which may extend to the years before its start.

I am intellectually indebted to Denny Borsboom, Angela Djupsjöbacka, Fabian Pfeffer, Anders Stenberg and many others. I also specifically want to thank my adviser Markus Jäntti, the two external examiners Roope Uusitalo and John Micklewright, custos Johan Willner and opponent Oskar Nordström Skans.

I gratefully acknowledge financial support from Yrjö Jahnssonin säätiö, Stiftelsens för Åbo Akademi forskningsinstitut, Bröderna Lars och Ernst Krogius forskningsfond, Jubileumsfonden and from the Academy of Finland.

Parts of this thesis were written at the Swedish Institute for Social Research. I appreciate the patience you have shown with me.

Kristian Koerselman
Helsinki, 2011

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The original question underlying this thesis concerns the effects of a particular educational policy: curriculum tracking. Curriculum tracking is the practice of stratifying students into educational tracks according to ability or achievement. Students usually follow a common comprehensive program up to a certain age, after which they are split up. The age at which the students are split varies from 9 or 10 years in some countries, to 16 or arguably 19 years in others.

Many authors have looked at curriculum tracking. Surveying the earlier literature, it is probable that tracking has larger (if such things can be compared) and more certain effects on the inequality of educational outcomes than on their average level. This is related to the literature on peer effects, where the achievement of any student depends positively on the achievement of

classmates. Tracking explicitly affects class composition, making classes more homogeneous and diminishing the magnitude of inequality-reducing peer effects.

The track chosen or selected into plays an important role in determining educational achievement and attainment and hence also occupation and earnings. Not only do higher tracks give access to better peer groups, it is often hard to change tracks at a later age. Once the student enters a lower track, the door to higher education may be all but closed.

Tracking should also be expected to have an effect on achievement *before* the start of tracking. Students, parents, teachers and principals know that the start of tracking is an important point in the student's educational and professional career. The students have an incentive to work harder before the tracking point, and parents, teachers and principals are likely to push them to do so.

Additionally, there is an incentive to substitute effort in non-tested subjects and proficiencies towards tested ones, crowding out non-tested subjects. Also, early tracking policies may be correlated with a more general competition-oriented educational system, perhaps also exposing students to tests at an earlier age.

For all these reasons, we should expect higher early age test scores in countries and regions that track early. One contribution of this thesis to the literature is to make a comprehensive empirical investigation of such incentive effects of tracking. I find evidence for incentive effects in both British and international data, at 0.09 UK standard deviations and 0.23 international

standard deviations respectively. The British estimate appears well-identified, and probably largely reflects the causal effect of tracking on early test scores. The international estimate should not be interpreted causally, as it probably reflects both a direct causal effect, and the effects of culture and institutions on both tracking policies and early test scores. The correlation is however remarkable, illustrating a strong connection between early tracking and early test scores. These findings can be placed in a larger literature that shows that students do indeed tend to respond to incentives.

I also look for incentive effects in the Swedish comprehensive school reform of the 1950s and 1960s. However, too many aspects of educational policy were changed at the same time, and it is hard to draw any substantive conclusions. Both positive and negative incentive effects are consistent with the data under different assumptions.

The existence of incentive effects is interesting in and of itself. How much competitive pressure students should be subject to at different ages is a matter under debate. Incentive effects have methodological implications as well. If pre-tracking scores are endogenous, we cannot use them to control for unobserved variables.

The second main contribution of this thesis is methodological. How well can we actually measure educational achievement? Economists use cognitive, noncognitive and achievement test scores as well as grades as if the numbers had cardinal meaning. In truth, we can only measure them at an ordinal level. This

makes the use of test score means, for example in mean-based regression such as OLS, questionable.

Fortunately, mean-based results seem to approximate underlying learning quite well, at least in an economic context and in well-behaved data. Nevertheless, economists should be aware of how test scores are constructed.

Both topics are highly relevant today. Tracking is a much debated policy in countries that still track early, such as for example Germany. At the same time, schools in late tracking countries such as Sweden are in danger of becoming more unequal through the introduction of voucher schools and other market based reforms. Tracking may be a viable meritocratic alternative to other forms of segregation.

The methodology behind educational achievement scores is more important now than ever. The public debate on education has become centered on educational achievement scores from international surveys such as PISA and TIMSS. It is important to know how these scores are produced, and what conclusions we can and cannot draw from them.

This thesis is structured as follows. In chapter 2, I go through some basic tools and concepts to analyze educational policy. In chapter 3 I explain why educational test scores are fundamentally measured at an ordinal level, and try to quantify how problematic this is in empirical work. In chapter 4, I go through the different forms of curriculum tracking, and selectively summarize the empirical literature on tracking. In chapter 5, I look at the incentive effects of tracking. Chapter 6 concludes.

Chapter 2

Tools and frameworks

2.1 The simplest possible framework

Following Hartog (2001), we can begin with "the simplest possible framework" to describe the educational process and subsequent labor market success.

$$\begin{aligned} s &= Ea \\ y &= p^{\mathrm{T}}Qs \end{aligned}$$

The educational production matrix E maps a vectors of abilities a into a vector of educational achievement or skills s. Skills are then mapped into wages y by subsequently multiplying skills by the productivity matrix Q and by a transposed vector of labor market prices p^{T}. Policy is understood to change E or $p^{\mathrm{T}}Q$, and

abilities to include educational inputs like parental 'quality'.

Throughout this thesis, I will assume that achievement has no value other than its effects on later outcomes. Wages are one (imperfect) measure of these later outcomes, but others are certainly possible.

There are two reasons for concentrating on later outcomes. On the one hand, I argue that viewing education as a means rather than an end fits the economic interpretation of education as an investment in personal productive capacity. This concept is called *human capital*, a set of skills and norms which increase productivity.

The concept of human capital was explored by William Petty as early as 1691, and has subsequently been studied by, among others, Smith, Engel, Walras and Fisher (Kiker 1968), with the most influential work on the subject probably being that of Schultz (1961), Becker (1964/1993) and Mincer (1974).

On the other hand, I argue that economic growth is one of the main concerns of policy makers. Even if it can be argued that the two reforms covered in this thesis were mainly redistributive in nature, productivity arguments played an important role in the internal and external debates preceding the reforms.

Due to data limitations, it is often impossible to estimate the effect of educational policy on wages directly.[1] Instead we have to contend ourselves with separate estimates of the two equations in the model. In the first step, we can for example look at the

[1] Among the exceptions are Meghir and Palme (2005) and Cunha et al. (2010).

effects of changing policy on s, and in the second at the effects of s on y.

There two potential issues with this approach. One is that policy can have effects on wages (or other outcomes) that are not mediated by measured achievement s. To some degree this problem can be ameliorated by testing achievement more broadly and thoroughly. To another we will have to accept the limitations that the data impose on us.

Another issue is that, depending on definitions, achievement is either an ordinal variable, or a cardinal variable which only can be measured ordinally. I will go through the implications of the ordinality of test scores further below.

2.2 Mapping achievement to wages

The mapping of educational achievement into wage outcomes has been investigated exhaustively. I summarize the literature in order to more easily evaluate changes in test scores in subsequent chapters.

On the micro level, the amount education that the individual chooses is usually modeled as an investment decision. Suppose that human capital loses all value at the end of the working life, but that it does not depreciate before that. Given a set of convenient other assumptions, it would be optimal for the individual to spend some part of his working life in education and the rest at work. The optimal amount invested depends on the marginal cost of education, an important part of which is the opportunity cost of not working, and the marginal discounted benefit of higher productivity later in life. Of course, since the persistent stream of education induced benefits only materializes after its competion, it pays to allocate all educational expenditures to the beginning of one's life.

We can estimate micro level returns to personal investment in human capital with a *Mincerian wage equation*, for example

$$\ln y = \alpha + \beta t + \gamma x + \delta x^2 + \varepsilon$$

where y are annual earnings, t denotes years of education and x years of labor market experience. The coefficient β is often interpreted as an approximation of the internal rate of return to education. In his 1974 book, Mincer estimates this particular

wage equation to be

$$\ln y = 6.20 + 0.11t + 0.08x - 0.01x^2 + \varepsilon$$

for white US men working in outside of the agricultural sector in 1959, implying that the rate of return of education for this group is about 11% (Mincer 1974). Newer studies with larger numbers of controls find estimates in the 6%–9% range (Ashenfelter et al. 1999).

It should be noted that the outcome of the educational production function is usually achievement, while a Mincerian wage equation is a function of the quantity of education. Achievement and quantity of education differ in two ways. First, quantity of education is an input measure while achievement is an output measure. Second, the quantity of education should be adjusted for quality. Some authors derive quality-adjusted measures of the quantity of education, but there is also a literature on the effects of changes in achievement on wages (e.g. Murnane et al. 2000, Altonji and Pierret 2001, Galindo-Rueda 2003, Lazear 2003, Speakman and Welch 2006, Hanushek and Zhang 2009).

The Mincerian wage equation is potentially problematic as a method to estimate returns. The highly educated may have higher unobserved abilities, which make them choose higher levels of education as well as contributing to their wages. This will bias the estimate of β upward. On the other hand, comparative advantages will bias estimates downward because without education, the highly educated would have made less than their peers, not more. Measurement error in educational attainment

will bias estimates downward as well.

Angrist and Krueger (1991) make an attempt to remove this bias by using natural variation in birth dates as an instrument for schooling: children born early in the calendar year have shorter schooling durations because compulsory schooling laws keep their younger peers in school longer. These IV results are not *much* larger than the regular, OLS results, suggesting that we can use normal Mincerian wage equations to approximate the true return to education. In fact, comparing 96 estimates from 27 studies, Ashenfelter et al. (1999) find surprisingly small differences between estimates from studies using all variation, studies using exogenous variation only, and twin studies.

A different interpretation is that both kinds of estimates are overestimates of the average return to schooling. Where OLS estimates under certain assumptions give an average effect for everyone, IV estimates only give an average effect for the 'switchers' who change their behavior because of the instrument. Because many of the instruments of choice mainly affect the disadvantaged, and because there is reason to believe that returns to schooling are larger for the more disadvantaged (Krueger and Lindahl 2001), this will inflate the IV estimates compared to the average causal effect.

Using a combination of instruments, it is possible to estimate marginal treatment effects at different parts of the distribution. Carneiro et al. (2005, 2010) do this, and estimate the shape of the distribution of the marginal returns to college attendance. They find decreasing, not increasing returns to college

for marginal college students, suggesting that average returns are higher.

Microeconomic estimates such as Mincerian wage equations estimate the private returns to education, but are uninformative of macro-level social gains or losses. In the typical OECD country, around 6% of GDP is used for formal education, five sixths of which comes from public sources (OECD 2010a, 2010b). How can such large subsidies be motivated?

There may be many different kinds of positive externalities associated with education. Milton Fiedman (1962, chapter VI) for example mentions stability and democracy, but believes that these externalities mainly arise from specific kinds of education. Thus we should be more willing to subsidize elementary schools and liberal arts colleges than vocational programs. Vocational education merely increases productivity, and individuals should be able to demand optimal amounts of it by themselves.

We may also think that education lowers crime rates and raises public health (Hartog and Van den Brink 2007, ch. 5) as well as labor market participation rates. There may be externalities in the invention and spread of new technologies, and of productivity in general.

On the other hand, education may have negative externalities as well. Perhaps education does not have an effect on productivity at all. In a world where education affects relative wages, but leaves average wages unaltered, the negative effect that any individual's education thus has on the wages of others can be seen as an externality. In such a world we would however observe a

positive regression coefficient in a Mincerian wage equation.

Suppose for example that employers cannot perfectly observe employee quality. If only highly competent employees can profitably *signal* their competence by obtaining costly advanced educational degrees, employers may be willing to offer higher wages to the educated, even when education adds nothing to that competence (Spence 1973). In such a setting, everyone would be better off if the general level of education were lowered while preserving its signaling value. Consequently society should try to discourage education rather than subsidize it.

Spence also points to relatively immutable but observable characteristics like sex or race, which can interact with the signaling variable. Some groups may be in one equilibrium while others are in another. Separate equilibria can exist because employers can observe the immutable characteristic. An employer may for example see a (lack of) education as a signal for men, but not for women.[2] This in turn removes the incentive for the women to educate themselves, sustaining the equilibrium.

Stiglitz (1975) compares pooling and separating signaling equilibria.[3] In the baseline case, education is costly, but both equilibria exist. Just like in Spence, society can get stuck in the suboptimal separating equilibrium where low quality workers have low wages, and high quality workers have to go through

[2] Or perhaps the other way around.

[3] Stiglitz speaks of 'screening', which is more of an objective assessment than the self-selective 'signaling'. However, in the context of education, they amount to largely the same thing. If we abstract educational degrees to a pass/fail indicator, and individuals are aware of their own abilities, they will choose a screening test which they are just able to pass, thus signaling their ability.

costly education in order not to be confused with low quality workers. Stiglitz then goes on to explore scenarios where signaling instruments can in fact yield gains to society. Among the examples are worker-job matching in case of comparative advantages, and the emergence of less efficient signaling devices in lieu of the current system.

Lange (2007) looks at wage dynamics in the National Longitudinal Survey of Youth. The result of the AFQT, a cognitive skills test, is available to the researcher, but not to the respondents' employers. Lange uses this to calculate the rate of employer learning about employee productivity. He is then able to put an upper bound of 25% on the signaling share of the returns to education.

Do the positive externalities of education outweigh the negative ones? There is some microeconomic evidence to support positive externalities. For example, Acemoglu and Angrist (2000) and Moretti (2004) find positive monetary external effects of a magnitude smaller than the private returns. On the macroeconomic level the evidence is rather weak (Krueger and Lindahl 2001, cf. Topel 1999).

Externalities per se seem to be an insufficient reason for the extensive government subsidies for education. A stronger argument can perhaps be found in concerns over the education of particular groups. Children of the lower classes may underinvest in education in a fully private system, either because of credit constraints, or because of cultural factors.

All these arguments regard the optimality of investment deci-

sions from an individual perspective. Apart from ensuring efficiency, educational policy can however also be used for social policy. Indeed, 'social' arguments for the large increase in public investments in education during the 20th century played as large a role as 'economical' ones at the time. An educational system can produce aggregate outcomes that society values, even if the size and allocation of educational investments can be inefficient from an individual perspective.

Chapter 3

Admissible statistics

Measures of educational outcomes are ordinal. In theory, any monotonic transformation of a test score distribution is also a valid distribution. One illustration of this fact can be found in figure 3.1, where I have collected histograms of nine test score distributions.

The distributions differ in skew, but this is not informative of skew in underlying achievement. In fact, we could administer two tests designed to measure the same achievement dimension to the same students, and obtain score distributions with different shapes.

When mapping test scores into wages, as in section 2.2, we can be pragmatic about this problem. Whatever the distribution of test scores we have, we can adapt the functional form of our model to make it fit the data. Whatever the true meaning of

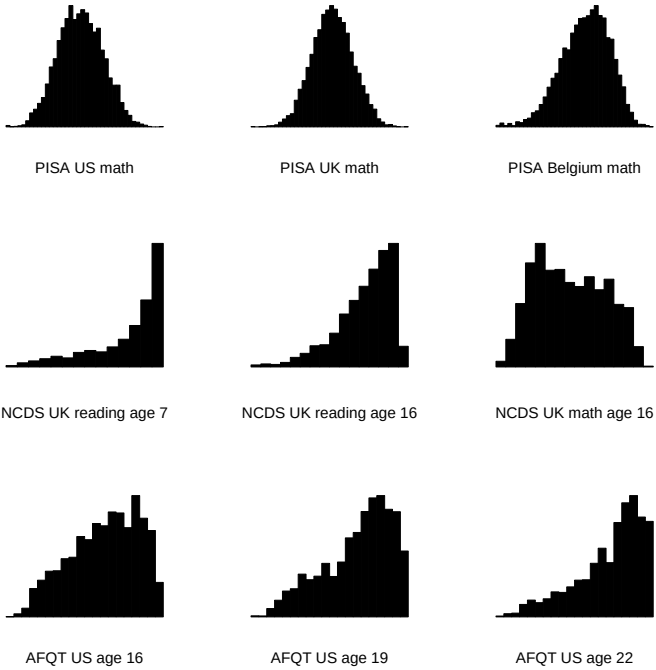

Figure 3.1: Test score distributions. Top row: cohort-representative weighted country score distributions for PISA 2006 math scores (OECD 2006). Middle row: reading and math scores for the NCDS (University of London 2008). Bottom row: weighted AFQT scores derived from the ASVAB administered to the NLSY79 (U.S. Bureau of Labor Statistics 2010) sample.

test scores, we can estimate the effect of a marginally higher score in any part of the test score distribution.

The first step in the framework on page 5 is more problematic. It describes the link between ability, background variables and policy on achievement and skills. This link is commonly estimated by regressing measures of ability or achievement on covariates, for example to evaluate educational policies or teacher performance (cf. Lazear 2003, Todd and Wolpin 2003, Hanushek 2006). A positive and monotonic transformation of test scores contains the same ordinal information as the original variable, but regression coefficients can change depending on if we use the original or the transformed variable. In extreme cases, coefficients can change sign.

The ordinality of test scores has not stopped economists from using and interpreting them as if they were cardinal. In this chapter, I summarize the psychometric theories behind test scores, and investigate how large robustness problems are in reality.

I find that the ordinality of test scores is not a large problem when treatment effects are homogeneous, or when we can also observe later outcomes to anchor the test scores to. Often, neither of these conditions are satisfied, and robustness problems can be large for arbitrary transformations of the test scores.

As economists, we can however interpret test scores as a measure of human capital, and restrict the set of transformations to those yielding distributions of human capital we deem reasonable. I estimate an empirical distribution of human capital, and show that normally distributed test scores are a close enough proxy for

underlying human capital for regression results to be relatively robust.

More skewed or irregular score distributions, such as those in the bottom two rows of Figure 3.1 may have to be transformed before use. We should also keep in mind that it is uninformative to report regression coefficients in test score points, and use standard deviations, percentile ranks or correspondent later outcomes instead.

3.1 Psychometric theory

Psychometricians have a long tradition of linking appropriate statistical methods to different kinds of data. A key insight is that all data are in essence mappings of empirical phenomena onto some scale or another, and that the choice of scale is to a certain degree arbitrary.

We want our statements to be qualitatively robust to changes in the mapping from the empirical world onto the data scale. For example, we do not want our qualitative conclusions to change when we map height into meters instead of feet. A comparison of mean heights of adult men in England and France should yield the same qualitative result as to which nation is the tallest in either case. Comparing mean height is indeed robust as the empirically taller nation will always have the larger mean height. By contrast, conclusions based on the mean of an ordinal variable are not robust to the choice of scale. Consider *highest completed education*. Using 1 for *primary education*, 2 for *upper secondary education* and 3 for *tertiary education* may or may not give a different ordering of the English and French means compared to using 9 for *tertiary education* instead of 3, even if $(1, 2, 9)$ is just as good a representation of the ordinal levels as is $(1, 2, 3)$.

Stevens (1946) suggests a relatively easy way to determine when we will run into robustness problems of the above kind. We group scales into four levels: nominal, ordinal, interval and ratio, as can be seen from Table 3.1. We call a certain statistic *admissible* for a level of scale when empirical conclusions derived from it are robust to the use different scales within the level.

Scale	Mapping	Examples of variables	Examples of admissible statistics
Ratio (highest)	$x' = ax$	income, age	coefficient of variation
Interval	$x' = ax + b$	school grade (i.e. year), calendar date	mean, variance
Ordinal	$x' = f(x)$, $f()$ monotonically increasing	level of education, socioeconomic background	median, other quantiles
Nominal (lowest)	$x' = f(x)$, $f()$ gives a one-to-one relationship	gender, race, religion	mode

Table 3.1: Admissible statistics for four different measurement levels, adapted from Stevens (1946). Each measurement level inherits the admissible statistics from the levels below.

Statistics are always admissible on higher level scales than their own, and inadmissible on lower levels.

A related but distinct problem is that of meaningfulness. We calculate statistics on our data in order to learn something about the real, empirical world. Statements on the data which bear no relationship to the empirical world, are therefore not empirically *meaningful* (cf. Hand 2004, section 2.4.1). A statement like "*The mean completed education in England is 1.8.*" makes no sense because it is not a statement on education in England as much as on the mean of a vector of arbitrary numbers stored on our computer.

Meaningfulness and admissibility usually coincide, but there may be situations in which they do not (cf. Lord 1953, Zand

Scholten and Borsboom 2009). We could for example compare mean education in England and France, and conclude that they are significantly *different*: that the English and French samples are not likely to have been drawn from the same population. The existence of a difference of the calculated means is dependent on the coding of the variable, and thus not robust, nor is the mean the best way to quantify this difference, but the conclusion that the religious composition of the two countries differ is meaningful nevertheless.

Test scores are used as a measure of a variety of concepts. They are designed to capture variables like intelligence or ability, proficiency at a certain task, or learning. Below, I will refer to the underlying variable as 'achievement' even though the reasoning applies to other variables just as well.

Achievement cannot be observed directly, but must be estimated using some kind of framework. These can be divided into two broad categories.

The simpler of the two is called Classical test theory or CTT. In CTT, the test score is a linear transformation of the proportion of test items or questions answered correctly. This is the kind of scoring we perhaps remember from our own time in school.

CTT is based on a true score model

$$x = t + \varepsilon$$

where t is the true, underlying probability of the student answering questions correctly, and x is the observed proportion of questions actually answered correctly. The error ε arises because

the number of questions is limited, adding noise to the estimate. We use x as the estimate of t.

Test scores calculated using CTT are straightforward to interpret. The scores are estimates of the proportion of questions a student would be expected to answer correctly when given a similar test. Group averages of CTT scores also have a clear interpretation: they are the proportion of questions the group as a whole would be expected to answer correctly. We could thus conclude that CTT scores are of ratio level, and we would be right to do so, if there were just one possible relevant test.

The advantage of CTT is however at the same time its disadvantage. CTT provides a score given a particular level of questions. The score distance between two students is determined by the level of questions considered. If the questions are very hard, almost no question will be answered correctly, student scores will be massed against the lower 0% bound, and consequently, the score distribution will have right skew (see Figure 3.2). Similarly, the score distribution will have left skew when the questions are very easy. In the first case, the score distances between low-scoring students become small, and between high-scoring students they become large. The opposite happens in the second case. (cf. Lord 1980, p. 50)

The difference in the skew of the score distribution affects our estimated mean effects. If we were to compare means between a treatment and a control group on the basis of the hard test, we would weight the right tail of the distribution more heavily, whereas if we were to compare means on the basis of the easy

test, we would weight the left tail more.

We can interpret CTT scores on a ratio level when speaking about a specific test. We could for example use a change in achievement test scores to identify that the average probability of answering correctly on a specific level of questions has increased. We cannot, however, generalize the result to the scores obtained by a different achievement test, even if both tests are designed to measure the same underlying concept. Also, we cannot make ratio-level statements on the effect on the underlying concept itself. We cannot conclude that 'mean achievement' has increased even if mean test scores have.

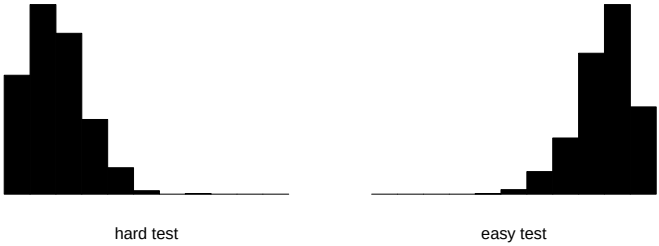

Figure 3.2: Hard CTT tests produce a score distribution with right skew while easy tests produce left skew.

An alternative to CTT is Item response theory, or IRT. IRT simultaneously estimates student and question properties by fitting an *item response function* which describes the probability of giving a specific response or answer to the specific item or question the function refers to. Often, the response categories are

'right' and 'wrong', and we estimate an item response function of the form

$$\mathbb{P}(y_{ij} = 1) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}}$$

This function is illustrated in Figure 3.3.

$\mathbb{P}(y_{ij} = 1)$ is the probability of student i answering question j correctly, θ_i is the level of student achievement we are interested in, and b_j is overall question difficulty. The inflexion point of the response function lies at $b_j = \theta_i$, and we say that student achievement and question difficulty are equal at this point.

The limiting probability of answering the question correctly for extremely low levels of achievement is given by c_j. It can be interpreted as the probability of answering correctly when making an uninformed guess. The upper probability limit is assumed to be one.

Question discrimination a_j determines the rate at which students get better at answering the question when they have a higher level of achievement. A question which everyone answers equally well has zero discrimination. a_j may even turn negative if answering correctly on question j correlates negatively with answering correctly on the others. This may for example be the case for trick questions. Questions with low or negative discrimination are however regularly discarded from test databases.

There are model variations where one or more item parameters are fixed or otherwise restricted, mainly because there is relatively little information contained in each response. When c is set to zero, and a to one, we obtain the commonly used Rasch

model. As is generally the case when $c = 0$, the inflexion point $b_j = \theta_i$ then lies at the level where the student is expected to answer the question correctly with probability 0.5.

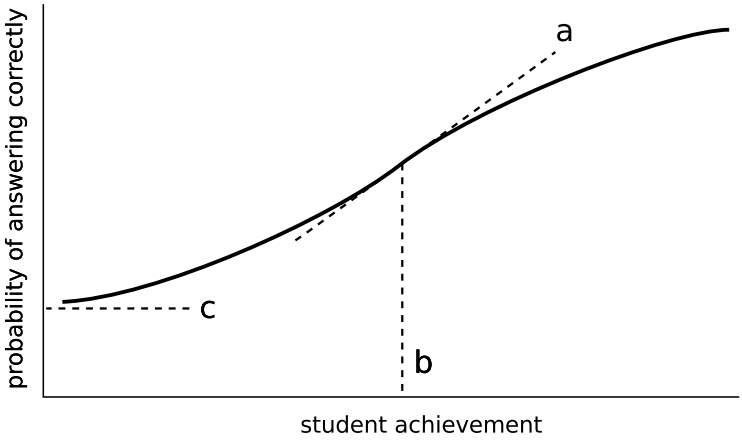

Figure 3.3: An item response function gives the probability of a student answering a certain question correctly as a function of his achievement. Question parameters a (discrimination), b (difficulty) and c (guessing) are illustrated in the figure.

IRT models remove much of the CTT models' dependence on question difficulty. Score distances arise from the difficulty with which students answer questions above and below their own level. If a student is answering questions above his own level of achievement with relative ease, the distance $\theta_i - b_j$ must be relatively close to zero, just as when he does not do unusually well on questions below his level.

Because similar scores can be estimated for question sets of different difficulties, it would seem that IRT models solve the CTT

models' ordinality problem. Unfortunately it returns in a different guise.

Unlike CTT-scores, IRT student scores are not anchored to some absolute measure. We can for example add a constant to the vectors θ and b and arrive at the same model fit. In the same way, we could multiply θ and b with a constant and divide a by it. The model is therefore unidentified if we do not impose additional restrictions on the scores, for example by specifying that the sample mean score equals zero, and its standard deviation one.

We can also change the higher moments of the distribution by estimating a different functional form. The horizontal achievement and difficulty axis can for example be transformed by $\theta^* = k_1 e^{k_2 \theta}$, where k_1 and k_2 are constants, so that both the item response functions and the distribution of θ and b are stretched out in one tail and compressed in the other (Lord 1980, p. 85), lending the score distribution arbitrary skew.

While CTT distances are a product of the particular test taken, IRT distances depend on the equations which we use to map raw scores into test scores. In both cases, we can reasonably change our methods, and obtain a test score which follows a different distribution. We must conclude that test scores are measured at an ordinal level, and that monotonic transformations of test scores are just as valid scores.

To illustrate how such transformations of test scores can change mean-based estimates, consider Figure 3.4. We compare a treatment (dashed lines) and a control distribution. Treatment has

a positive effect in the right tail, but a negative effect in the left. In the original data (top), the positive effect outweighs the negative, and the mean treatment effect is positive. In the normalized version of the data however (bottom), the right tail is given less weight, and the mean treatment effect turns negative. The estimated treatment effect is not robust to a monotonic transformation of the test scores.

As can be seen from Figure 3.5, ordinal methods such as quantile regression are qualitatively robust in the sense that for each quantile, the estimated effect has the same sign both in the true and in the normalized distribution.

The ordinality of test scores is most elegantly handled by using statistics of the ordinal level. A statement like 'the median student in the treatment distribution has higher achievement than the median student in the control distribution' needs no cardinal interpretation, and quantile-based methods are qualitatively robust. Psychometric theory would discourage us from using mean-based analysis as it is based on information we cannot actually measure.

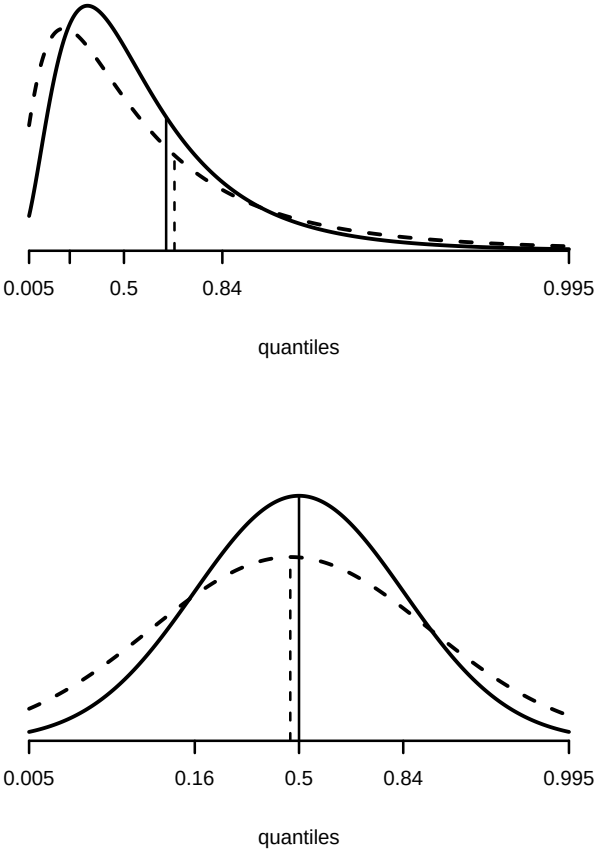

Figure 3.4: If the true distribution (top) differs from the imposed one (bottom), this may lead to qualitatively wrong conclusions when comparing distribution means. In this case, the treatment distribution (dashed lines) has a higher mean in the true data, but appears to have a lower mean after normalization.

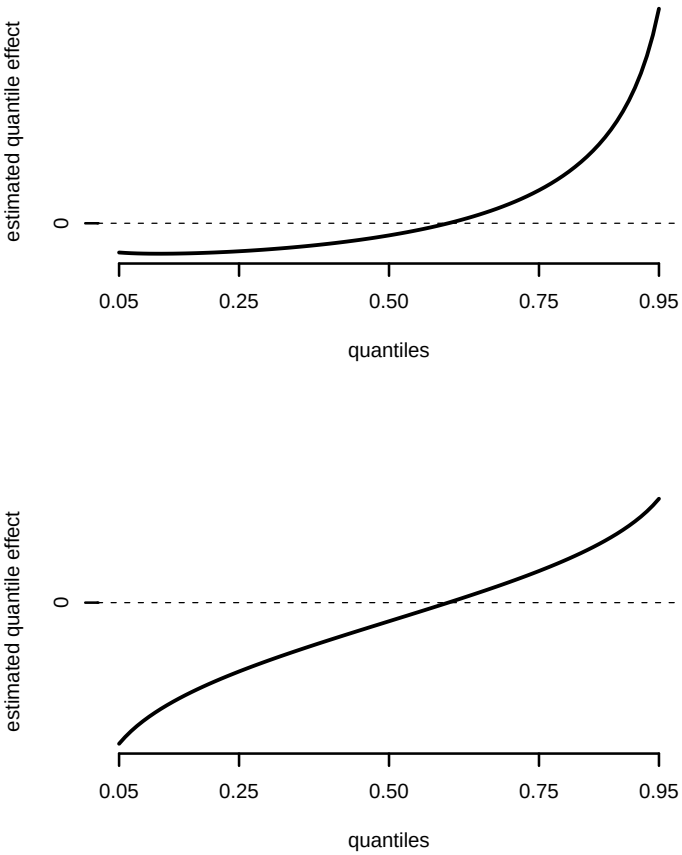

Figure 3.5: Quantile-based methods are qualitatively robust, with a negative effect on all quantiles below about 0.6, and a positive effect on all quantiles above for the treatment effect from Figure 3.4.

3.2 Economic practice

Due to the problems outlined in section 2.1, it is common practice in the economics of education literature to regress test scores themselves on other variables (cf. Lazear 2003, Todd and Wolpin 2003, Hanushek 2006). As most regression techniques (such as OLS) are mean based, this raises questions both of interpretation and of robustness.

Quantitative statements like "mean achievement has increased by 11 points" are hard to interpret because there may not exist such an empirical concept as mean achievement. Is it as meaningless to compare means of achievement as it is to state that the mean completed education in England is 1.8?

To a psychometrician, a comparison of mean test scores may indeed be meaningless. Comparisons of means are only meaningful if achievement distances are comparable across the distribution: is a ten point increase by one individual the same thing as a ten point increase by another, in a different part of the distribution? If achievement is truly ordinal, point distances are arbitrary, and clearly not comparable.

The economist however, can interpret test scores as a measure of human capital as per the framework in chapter 2 (e.g. Hanushek 2006, section 2). Human capital can be valued at market prices, i.e. wages, and the price of human capital is clearly of ratio level. A statement like "Adam has twice as much human capital as Bert" is much less problematic than "Adam is twice as intelligent as Bert."

Treating achievement as human capital also diminishes problems of measurement. Of course, the existence of an underlying cardinal concept of achievement does not mean that we can measure it at that level. We can however use theory as well as empirical information from other data sets to impose a distribution onto the measured scores. This is something we are already doing implicitly by treating the test scores we have as cardinal. It seems that we could easily improve on this method by explicitly considering whether the cardinal information we use is representative of underlying human capital.

The human capital interpretation of achievement also implies that we do not need to account for robustness to arbitrary monotonic transformations, but only to transformations that yield reasonable human capital distributions.

Which possible underlying distributions would seem reasonable? The normal distribution stands out as a natural candidate. It has theoretical appeal as it emerges from an addition of many independent draws from an arbitrary, finite distribution per the central limit theorem. If we think of learning as an additive random process, where each day's new learning is a random draw to be added to the existing stock, achievement will be normal.

A second appealing distribution is the the lognormal. It has a similar relationship to the central limit theorem as the normal: if we multiply rather than add the (positive) draws, we will end up with a lognormal distribution. We can think of learning as a process in which students start from the same baseline, and learn at small, random rates each day, finally arriving at their test-

day achievement level. The drawing of learning rates rather than amounts implies that we think that past and future learning is correlated on an absolute level, so that higher achieving students are expected to acquire more additional knowledge in the future than their peers.

There are other reasons which make the lognormal distribution appealing. Even if learning would be additive in principle, the achievement distribution will have right skew if high ability individuals also put more effort, time or other resources into learning (cf. Becker 1964/1993, p. 100).

The standard assumption in the economics of education literature is that test scores map exponentially into wage income (cf. Lazear 2003, Hanushek 2006, Cunha et al. 2010). Given that test scores are often normally distributed, this implies that the distribution of human capital is lognormal when measured in wages.

Suppose that we regress normally distributed test scores, would results be very different if we were to use a lognormal human capital distribution instead? To try to answer this question, I compare regression results between using normal test scores, and their lognormally distributed empirical monetary value. In order to make the difference between the two distributions as large as reasonably possible, I try to arrive at an overestimate of the true return to achievement

I use the UK National Child Development Study (NCDS 2008). I select males who are full-time employed at age 48, and calculate the principal component of their normalized age 11 and 16 test

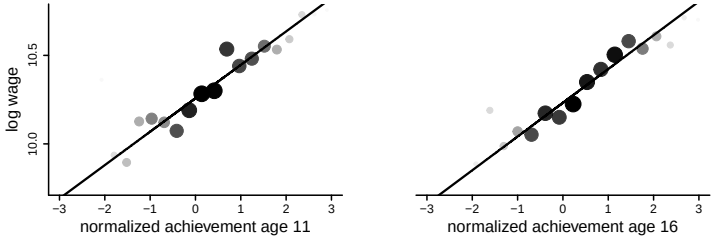

Figure 3.6: Average logged gross wages of 48-year old full-time employed males for different achievement levels (circles, circle area and color is proportionate to the number of observations) and the regression line through the unaveraged data. Data: NCDS 2008.

scores. The use of a wage measure at a relatively old age should increase the size of the estimate because the impact of test scores should rise with age (Altonji and Pierret 2001, Galindo-Rueda 2003, Lange 2007).

Figure 3.6 shows average logged age 48 gross wages for different achievement intervals at age 11 and 16 (circles). In line with the economic literature, there is indeed an approximately linear relationship between test scores and logged wages, which implies an exponential mapping from scores to wages.

Part of the relationship between achievement and wages is due to selection, and part is due to the causal effect of achievement on wages. We can explore this relationship by estimating variations of

$$\ln w = a + by + Xc + \varepsilon$$

where w are wages, y is the first principal component of either age 11 or age 16 normalized achievement scores, and X is a matrix containing control variables.

Because y is a noisy measure of underlying achievement, this kind of specification will create an attenuated estimate of the true relationship between achievement and wages. Luckily we have multiple measures of achievement for both ages. Under the assumption that the measurement error is white noise, it is possible to correct for this bias by rescaling the measures with their respective reliability ratios of the measure in each specification (cf. Griliches 1986).

The reliability ratio is given by the ratio of the variance of the latent variable or signal to the variance of the measure:

$$\frac{\sigma^2_{signal}}{\sigma^2_{signal} + \sigma^2_{noise}}.$$

It can be estimated by comparing measures of the underlying variable with other measures of the same variable. For example, if they are perfectly correlated, the reliability ratio is one. If they are completely uncorrelated, it is zero.

I rescale the principal components so that the signal contained in them has a standard deviation of one, and we can interpret regression results as the increase in log wages associated with a one standard deviation increase in *underlying* achievement.

I start by regressing log wages on achievement only. The estimated values of b can be found in the first column of Table 3.2. They are moderately large at wage differences of 0.21 logs for a

dependent variable: log wage	(1)	(2)	(3)
controlling for measurement error:			
age 11 score	0.215	0.189	0.134
age 16 score	0.235	0.212	0.156
not controlling for measurement error:			
age 11 score	0.189	0.164	0.112
age 16 score	0.191	0.167	0.112
controls:			
parental background		yes	yes
highest educational attainment			yes

Table 3.2: Separate estimates of the relationship between either age 11 or age 16 standardized test scores and logged age 48 wages for full-time employed males in the British 1958 cohort. Data: NCDS (2008).

one standard deviation increase in age 11 test scores, and 0.23 for age 16 scores.

In column (2) I add controls for socioeconomic background. I arrive at an association between test scores and log wages of 0.19 for the age 11 achievement distribution and 0.21 for age 16.

Depending on what we want to condition the wage distribution on, we can add more controls. Column (3) shows the estimates when I control for (endogenous) educational attainment (standardized to ISCED levels) as well. This reduces the estimates to 0.13 for age 11 and 0.16 for age 16.

Other authors have found estimates in the range of 0.00–0.20 for all three kinds of specifications, with higher values for the US than for other countries (Murnane et al. 2000, Altonji and Pierret 2001, Galindo-Rueda 2003, Lazear 2003, Speakman and

Welch 2006, Hanushek and Woessmann 2009, Hanushek and Zhang 2009). Compared to these, my estimates are indeed at the high end of the range. As can be seen from the bottom six estimates in the table, this is in part because I control for measurement error in the test scores, but probably also because I purposefully do not adequately control for selection and the use of late-age wages.

The next step is to map the standardized score distribution y into the estimated conditional wage distribution $\hat{w}$ according to

$$\hat{w} = \exp(\hat{a} + \hat{b}y)$$

The conditional wages $\hat{w}$ will be distributed lognormally with the distribution parameter σ equal to $\hat{b}$. The larger σ, the larger the skew of the lognormal. For low values of σ, the conditional wage distribution will look like the normal. The largest value of b in Table 3.2 which we still might reasonably call causal stands at 0.212 in column (2), even though the true causal effect is probably smaller. The lognormal distribution we obtain by setting σ equal to 0.212 can be seen from the right panel in Figure 3.7.

Suppose that the true cardinal achievement distribution is given by this conditional wage distribution, but that we use a normal test score distribution for our analysis. How much will regression results change?

To keep things simple, let us compare means between a treatment (subscript t) and a control group (subscript 0). I will call the difference between the two the treatment effect on the mean

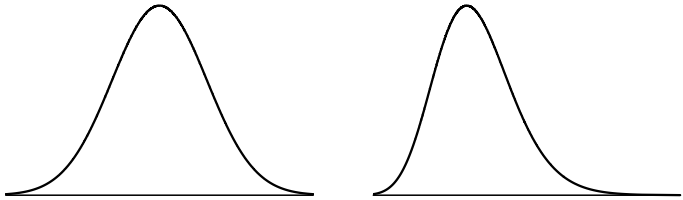

Figure 3.7: The estimated wage distribution conditional on differences in achievement levels, controlling for parental background. A one percentile in the achievement distribution (left) is associated with a one percentile increase in the age 48 wage distribution (right). Data: NCDS 2008.

β_μ. Suppose that the true distribution is lognormal, and given by

$$Y \sim \text{LN}(\mu, \sigma^2),$$

but that we measure normal data given by

$$y = \ln(Y) \sim \text{N}(\mu, \sigma^2).$$

In order to catch only the effect of a change in the shape of the distribution, and not the effect of a change in the scale, I will compare the difference of means in the normal distribution with the difference of logged means in the lognormal distribution. This implies that the difference will be expressed in terms of the normalized test scores.

The estimate of the difference between the means β_μ is biased by:

$$\text{bias} = (\mathbb{E}[y_t] - \mathbb{E}[y_0]) - (\ln(\mathbb{E}[Y_t]) - \ln(\mathbb{E}[Y_0]))\,.$$

In terms of the moments of the treatment and control distributions, this equals

$$\text{bias} = (\mu_t - \mu_0) - \left(\mu_t + \frac{1}{2}\sigma_t^2 - \mu_0 - \frac{1}{2}\sigma_0^2\right) = \frac{1}{2}\left(\sigma_0^2 - \sigma_t^2\right).$$

Let us define β_σ as the amount by which the treatment distribution is wider than the control distribution. Note that β_σ is expressed in control group standard deviations.

$$\sigma_t = (1 + \beta_\sigma)\sigma_0$$

Rewriting the earlier equation in terms of σ_0 and β_σ we then get

$$\text{bias} = -\sigma_0^2 \left(\beta_\sigma + \frac{1}{2}\beta_\sigma^2 \right).$$

This shows that the amount of bias generated by assuming a normal distribution where the lognormal distribution is appropriate is independent of the treatment effect on the mean, but dependent on the difference in variance between treatment and control groups. A relatively larger variance in the treatment group will lead to a negative bias in the estimate of the treatment effect, and vice versa.

The independence of qualitative robustness of the means of the distributions can be generalized. Davison and Sharma (1988) show that mean differences between two normal distributions of equal variance are indicative of mean differences in any monotonic transformation of those distributions.

How large is the bias in practice? In many cases, variances are more or less constant over treatment, and the bias will be close to zero in accordance with Davison and Sharma. One example where this is clearly not the case is curriculum tracking. Tracking almost certainly leads to larger differences between students. We can thus use curriculum tracking as a kind of worst-case scenario for educational policy analysis.

I have selected three recent empirical papers from the literature on the subject from which to get empirical values for both β_μ and β_σ. All three papers use approximately normally distributed

test scores.

Hanushek and Woessmann (2006) compare tracking policies between countries cross-sectionally on the basis of PISA/PIRLS and TIMSS data. Pekkarinen et al. (2009a) investigate the effect of the 1970s Finnish comprehensive school reform using panel data, while Duflo et al. (2008) use a randomized trial in Kenya to look at the effects of tracking. These are three quite different settings, and their respective results are not necessarily generalizable across regions and times. It is therefore perhaps not surprising that the three papers find significant effects on the mean of different signs. Tracking is however associated with larger differences between students in all three papers.

For Hanushek and Woessmann, I have taken the pooled estimates from Tables 3 and 4 directly. For Pekkarinen et al. I use the fourth column in Table 4 as well as the descriptive information in Table 3 to approximate the standardized effects on mean and standard deviation. For Duflo et al. I use the second column in Table 2 for the effect on the mean directly, and the heterogeneous estimates in column 4 of the same table to approximate the effect on the standard deviation. In all three papers, test scores are approximately normally distributed.

The first column in Table 3.3 shows standardized estimated treatment effects on the mean from these papers. The second column contains the standardized effects on the distributions' standard deviations. The third column contains the parameter σ_0, which determines the skew of the assumed underlying human capital distribution. I assume the human capital distribution to

	(1)	(2)	(3)	(4)	(5)
Paper	β_μ	β_σ	σ_0	bias	new β_μ
Hanushek and Woessmann	-0.179	0.101	0.212	-0.005	-0.174
Pekkarinen et al.	-0.010	0.009	0.212	0.000	-0.009
Duflo et al.	0.175	0.042	0.212	-0.002	0.177

Table 3.3: Estimated treatment effects of curriculum from a number of selected papers. The last column shows the treatment effect on the mean under the assumed lognormal achievement distribution.

have a σ_0 equal to the 0.212 estimated above.

The fourth column contains the size of the bias under the assumption that the true distribution is lognormal, and the fifth the updated mean treatment effect. The size of the bias is small: not exceeding 0.005 of a standard deviation in test scores for any of the papers. This is not enough to change the papers' respective quantitative conclusions very much, and will certainly not change the sign of the estimates.

The estimate $\hat{b}$ underlying the conditional wage distribution used for this analysis is probably an overestimate of the causal relationship between test scores and wages. Were we to use a smaller value of $\hat{b}$, the corresponding biases would be smaller in size as well.

In conclusion, we can see that even if the policy evaluated here has considerable effects on the spread of the test score distribution, estimates of effect sizes vary very little whether we assume that the latent trait is normally or lognormally distributed when we give the lognormal distribution reasonable parameters.

3.3 Discussion

In part driven by limited availability of data on later outcomes, economists often regress test scores on policy, teacher and background variables. This is potentially problematic because test scores are fundamentally ordinal measurements.

This chapter illustrates that normally distributed test scores are a reasonable approximation of underlying human capital. In cases where later outcomes are not available to the researcher, the use of mean-based methods on normal test scores seems reasonable.

This result does not necessarily generalize to more exotic test score distributions. In cases where score distributions are irregular or skewed, like in the bottom two rows of Figure 3.1, we may do better to normalize test scores before use, at least if we think that our sample is reasonably representative of the population.

It should be remembered that while we cannot truly measure the underlying variable on a higher level than the ordinal, the measurements are not entirely void of cardinal information either. If we estimate test scores with a procedure that is known to yield a reasonable cardinal measure for the population, using the same procedure on a select sample of that population will quite adequately recover the underlying cardinal variable for that sample, even if it follows a very different distribution.

In either case, the researcher should be conscious that score distances are dependent on the assumptions going into their estimation, and that the quality of his inference is dependent on

the quality of these assumptions.

I advocate the use of methods such as quantile regression on test scores, at the very least as a kind of robustness check. The bias produced by using the wrong distribution of test scores increases in the heterogeneity of the estimated effect. Those cases where mean-based analysis is the least robust will thus also be the cases where quantile regression will give the most interesting results, worth reporting in and of themselves.

It is hard to compare the size of regression coefficients between analyses that use different sets of test scores. If treatment leads to a six point increase in one test, and a different treatment leads to an eight point increase in a different test, it is impossible even to guess which treatment has the larger effect without additional information. One way to partly circumvent the problem is to standardize test scores and use a metric relative to the sample standard deviation. This is done explicitly in international surveys like PISA.

When standardizing test scores, e.g. when reporting effect sizes, we should however keep in mind that measurement error in the test scores will affect the standardization, even if the variable is used as a dependent variable only. The effect size is attenuated with the square root of the reliability ratio.

Consider for example a policy that has a one standard deviation effect on underlying achievement. If we have a single measure of that achievement with a (not unusually small) reliability ratio of 0.5, the measured effect on the standardized achievement measure will only be 0.7 standard deviations. Estimates based on

a noisy standardized test score variable will thus be attenuated relative to the underlying variable even if the test score is the dependent variable in the analysis.

In my own empirical analyses below, I will use mean based methods on test scores, both out of necessity and out of conviction.

On the one hand, the methods I use have been developed for comparing means, and quantile based estimators are not as widely available. On the other, I make sure to consistently use normal or normalized test scores. I also note that estimated effects are reasonably homogeneous in each case. The effect on the mean can therefore be seen as a good approximation of the effect on the median.

Chapter 4

Curriculum tracking

I now turn to actual educational policy evaluation. Before starting with the main empirical contribution of this thesis however, I summarize parts of the literature relevant to my work.

4.1 Peer effects

A student's achievement is not only dependent on his own ability or socioeconomic background. Equally important for his achievement are his classmates. The influences of other students are usually called peer effects. In their simplest form, peer effects cause a student's achievement to regress to the mean of his classmates' achievement.

Peer effects can work through various mechanisms which can be hard to disentangle quantitatively. For example, peer effects

may change student characteristics such as ambitions directly, but may also work more indirectly through class culture and teaching styles, or even more indirectly through interactions between parents and teachers or principals.

Direct peer effects are likely be more reflective or multiplicative than more indirect ones, so that a positive shock to one member of the peer group causes improvements in each of the other members, and so on until a new equilibrium is reached.

The existence of peer effects suggest that one relatively straightforward way to change inequality in educational achievement is to change class composition. As will become clear further below, this is indeed the case.

In theory, peer effects can take many forms. In the standard, *linear* model, the peer effect on each student is a linear function of the average achievement of his peers. On either side of the linear case are peer effects *convex* or *concave* in means. Where there is no effect of sorting on average outcomes in the linear case, convex peer effects imply that sorting is efficient, while concave peer effects imply that it is inefficient.

In the *bad apple* and the *shining light* models, only the worst respectively the best student matter for the peer effect. The bad apple model implies that sorting is efficient, the shining light model that it is inefficient.

In the *boutique* and *focus* models, peer homogeneity aids educational production. In the boutique model, it is important that there are other students similar to each single student. In the focus model, homogeneity is also good if the single student is

not part of the homogeneous group, for example because teachers respond badly to heterogeneous classes. The *rainbow* model is the opposite of the focus model: students need other kinds of students around them.

In the *invidious comparison* model, individuals care about their rank, and are discouraged by peers who do better than themselves (and vice versa).

Model	Description	Sorting efficient?
Linear	Uniform positive effect of mean peer composition.	+
Convex	Positive and increasing effect of mean peer composition.	+
Concave	Positive but decreasing effect of mean peer composition.	–
Bad apple	Only worst student matters.	+
Shining light	Only best student matters.	–
Boutique	Important that there are no isolated students.	+
Focus	Positive effect of class homogeneity.	+
Rainbow	Positive effect of class heterogeneity.	+
Invidious comparison	Relative ranking positive effect for single student.	(?)

Table 4.1: Possible types of peer effects, loosely after Hoxby and Weingarth (2005).

The peer effects model of Lazear (2001) is often used as a building block in larger models. In the Lazear model, students are thought to have an individual probability to disrupt teaching $(1 - p_i)$, either by causing disturbances or by asking questions all other students already know the answer to. A class of size n

is thus being educated at any given moment with probability

$$\mathbb{P} = \prod_{i=1}^{n} p_i$$

This particular functional form makes peer effects convex, and grouping students efficient. If all classes are to have the same size, it can easily be seen that the presence of stratified classes will increase average achievement, as well as differences between students. When allowing class size to differ, taking into account a budget constraint, and solving for optimal class size, higher tracks, with high values of p, should have bigger classes than lower tracks. Groups of students with sufficiently low values of p should be separated from other students, and put into large classes where disturbances approach 100%.

Though the model is elegant, a number of objections should be raised against this model. First, its results follow directly from the assumed convexity of peer effects. Convex peer effects are however not strongly supported by the empirical literature. Moreover, even if the mechanism seems plausible at first sight, it is assumed that individual values of p_i are not themselves affected by peers. If they are endogenous, peer effects may be linear or concave nevertheless.

Empirical results on peer effects are mixed.

Hoxby (2000) estimates linear peer effects in a large sample of panel data from about 3300 Texan schools by using naturally occurring, unexpected year-to-year changes in class composition as an exogenous source of variation. This results in an estimated

positive peer effect of between 0.15 and 0.4 points for every one point change in peers' reading scores.

Ammermüller and Pischke (2006) find that class placement within primary schools are random within their sample and estimate a peer effect of 0.11. They find little evidence that peer effects be nonlinear.

Hoxby and Weingarth (2005) try to reject some of the peer effect models discussed above. While their results are not conclusive, they find evidence for a combination of linear peer effects and a return to homogeneity as in the boutique and focus models.

Hoxby (2000) also estimates peer effects for gender and for different ethnic backgrounds. Interestingly enough, girls have a positive peer effect on both boys' and other girls' math scores. The effect is so large that it is unlikely to be caused by a gender-neutral mechanism, i.e. by girls' higher overall achievement alone. Probably, the benign effect of girl to boy ratio is indirect, for example through class culture. There is some evidence that stronger peer effects exist *within* ethnic groups.

Lavy and Schlosser (2007) use a similar method to estimate gender-specific peer effects. Girls have a positive effect on their peers, and the effect is stronger at higher proportions of girls. The effect even exists for subjects where boys do better. A questionnaire shows that the positive effect of girls is partly due to an improved in-class learning environment.

Lavy et al. (2009) find that peer effects mainly originate in the top and bottom 5% of students, and that it is ability, not parental background, that causes the peer effect. Top students

have a positive effect on girls, and a negative effect on boys, while bottom students are bad for both sexes.

To the degree that the peer effects literature seeks an answer to which allocation of students to classes or schools is most efficient, I feel that it is partially misguided. As I have discussed in chapter 3, the distributional form of educational achievement is to a certain degree arbitrary. Changing the assumed achievement distribution will change the implicit weights given to the achievement of students in different parts of the distribution, thus changing efficiency. For example, if peer effects are linear in normalized test scores, they will look convex in a test score distribution with right skew, and concave in a distribution with right skew. If sorting is neutral in the first case, it will look efficient in the second, and inefficient in the third.

Peer effects are however also relevant for distributional concerns, as well as for the political economy of private schooling and segregation. In such models, peer effects are thought to be positive, but their convexity or concavity matters less.

If educational achievement were the only relevant variable, peer effects combined with a system in which coalitions of students are allowed to set up their own schools would lead to an equilibrium in which schools are fully stratified by ability. In such an equilibrium, no school would accept a student with an ability under the mean ability of the school, and no student would accept a school with a lower mean ability than himself. Hansmann (1999) argues that such stratification reduces competition between universities in the US.

If educational institutions can set tuition fees freely, and students care about both achievement and money, peer effects can facilitate trade between achievement and money. A rich, low ability student can pay a poor, high ability student. In return, the high ability student gives up the positive peer effects that a cognitively homogeneous class at his level would offer, and shares a class with the rich, low ability student instead. If we see a student's ability and his parents' wealth as belonging to him, an argument can be made that both efficiency and fairness require that trade be allowed in the two educational inputs.

Trade between wealth and ability will lead to a hierarchy of schools, with increasingly high tuition fees and increasingly large scholarships to attract able students. The top panel of Figure 4.1 shows a partitioning of student space along wealth and ability. The public school at the bottom left has the lowest tuition fees and the lowest peer group quality, while private school 4 in the top right corner has the highest.

Epple and Romano (1998) present a model for exactly such a market, and discuss the effects of vouchers in a system which has both exclusive private, and publicly funded schools. Private schools charge tuition according to ability, while public schools are free and open to all. Trade in inputs is initially restricted because of the implicit subsidy to the public schools. As vouchers are introduced, and students get more freedom to self-organize, private schools become cheaper. Unsurprisingly perhaps, the voucher system hurts poor low ability students who see their better-quality peers leave the public school system. Epple et al. (2004) find evidence for the relevance of this model in the US

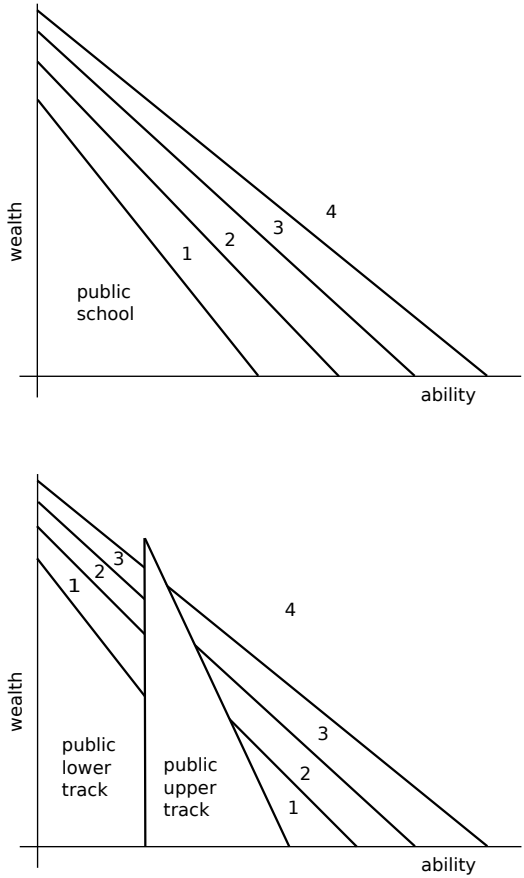

Figure 4.1: A market for education with one public school and private schools 1 to 4 (top). A public school sorted according to ability (bottom) can induce high ability students to enter the public system, reducing the amount of stratification on wealth. Adapted from Epple and Romano (1998) and Epple et al. (2002).

market for education.

If we dislike the outcome sketched up above because we feel that income or wealth based educational stratification is wrong, it may be possible to counter such stratification by tracking students based on ability in the public schools (Epple et al. 2002). This is illustrated in the bottom panel of Figure 4.1. The tuition-free upper public track attracts high ability students that prefer a private school in the untracked system.

In a similar fashion, Figlio and Page (2000) argue that schools in disadvantaged neighborhoods can attract high income students by tracking their students. This may be advantageous even to the lower track students if there are school level peer effects across the tracks or if sharing a school with high income high income students has other advantages.

The dynamics sketched up here suggest that public school tracking can be a second-best solution when a small private sector is desirable, but mandating a public monopoly is politically unfeasible.

There may be other ways to achieve a similar outcome. Chakrabarti (2005) develops a model of a system where voucher schools are not allowed to select between applicants. Schools do nevertheless not become completely destratified by ability because only committed parents send their children to the voucher schools. When schools are not allowed to charge fees on top of the vouchers they receive, the system does however prevent sorting by income.

She then looks at the Milwaukee voucher program, which re-

quires random selection among applicants and forbids schools to charge additional fees, and does indeed find sorting by ability but not by income.

4.2 Dimensions of tracking

Curriculum tracking is the explicit separation of students into schools or classes based on observed past or expected future achievement. While it is uncommon to explicitly track at the primary level, and the norm is to do so at the tertiary, there are large differences in tracking policies at the secondary level. Since the Second World War, some countries have postponed tracking from the end of primary school to the end of middle school or even to the end of high school, while others have left their tracking policies unchanged (Benn and Chitty 1996, p. 7; Marklund 1980, ch. 13). This makes questions on the effects of tracking highly relevant. At the same time, the variation in tracking policies, both temporal and spatial, provides us a means to identify its effects.

Tracking policies vary in many dimensions. Perhaps the most visible dimension is the age of first tracking. Related measures are the number of years to be spent in tracked education and the number of tracks at a certain age. The first two measures attach an age to a certain threshold amount of tracking, the third assigns an amount of tracking to a certain reference age.

Tracking may be class-based or school-based. School-based tracking is more relevant in an European context, with much varia-

tion in school-based tracking policies between countries. International comparisons of school-based tracking policies will therefore tend to ignore strong within-school tracking in the US, Canada and elsewhere (Betts 2010). In my empirical analysis below, I will mainly concentrate on the age at which school-based tracking starts.

Another dimension of tracking is whether stratification is horizontal or vertical. Vertically stratified tracking systems create a hierarchy of ability or achievement, while horizontal systems are based on a difference in subject matter. Some tracks may for example stress the arts more, while others stress scientific subjects. (cf. Sorensen 1970)

It is also possible to differentiate systems along dimensions like selectivity, electivity, inclusiveness, scope and track mobility (cf. Gamoran 1992). Selectivity is a measure of the difference between tracks, or of the homogeneity within them. Electivity refers to the extent to which students can choose their track placement themselves. Inclusiveness indicates the relative sizes of the tracks: when most students are placed in one track, it is said to be inclusive. Exclusive tracks create stigma or prestige, depending on which students end up in it.[1] Scope refers to the relative amount of classes that are taken inside the track. A school may for example track mathematics classes while keeping common language classes. Track mobility indicates how easy is it for the students to change tracks once they have entered one.

The multidimensionality of tracking policies brings empirical

[1] Gamoran defines inclusiveness as the relative size of the *upper* track.

problems with it. Combinations of policy changes may have non-additive effects on outcomes. Perhaps tracking early is not as bad for intergenerational mobility if the tracks are horizontally sorted, or if track mobility is high. Because new policies often entail changes to multiple tracking dimensions, their respective effects can be hard to disentangle. We will see an example of this in my analysis of the Swedish comprehensive school reform in section 5.3.

Nonexperimental comparisons on the other hand suffer from measurement problems. Internationally comparable information on tracking policies is scarce, especially outside of the European Union, and official policies may not reflect de facto segregation of students. (cf. Brunello and Checci 2007)

I have illustrated the problem in Tables 4.3 and 4.4, where I have collected similar tracking variables from six different papers on early tracking. Inclusion in the table is conditional on inclusion in my own analysis in chapter 5. My assessment based on tracking status at age 12 or 14 is listed in columns (K1) and (K2). Columns (H1) and (H2) give the same measure for thresholds 14 and 15 from Hanushek and Woessmann (2006). Waldinger (2006) uses an earlier threshold, listed in column (W). Pfeffer (2009) uses an early tracking indicator with three levels, listed in column (P). Bedard and Cho (2007) use the number of years spent in a comprehensive system (B1) and a measure based on the percentage of higher track students (B2). Brunello and Checchi (2007) use the age of first selection (C1), the percentage of primary and secondary education spent tracked (C2) and the share of students in upper secondary vocational (C3).

Ammermueller (A) uses the number of tracks (Ammermueller 2005).

In Figure 4.2 I have illustrated the correlations between these measures. As can be seen from both the tables and the figure, authors make substantially different tracking assessments even when using similar measures.

Dimension	Description
Age of split	Children usually attend comprehensive primary schools, to be split up into tracks at a certain age. The later this point is, the less stratified is the system overall.
Duration	Number of years students spend in tracked systems. Related to age of split.
Number of tracks	It can be argued that systems with a large number of tracks are more stratified than those that have only two.
Vertical or horizontal	Vertical tracking refers to a difference in levels, horizontal tracking to a difference in subjects.
Selectivity	Selective systems have larger differences between tracks, and are more stratified.
Electivity	In elective systems, students can choose themselves which track to enter. This is more usual in horizontally sorted tracks (Sorensen 1970)
Inclusiveness	The more students attend a certain track, the more inclusive it is. Exclusive tracks can create excessive stigma or prestige. A small but horizontally sorted track (such as for example a full-time music school) may not affect overall stratification very much.
Scope	Proportion of time students spend in the tracked system. A school may for example track on a single subject only. Tracking systems with less scope are less stratified.
Track mobility	How easy is it to switch tracks past the age of split? The effects of tracking on intergenerational mobility may be smaller the easier it is for the student to change his mind later on.

Table 4.2: Dimensions of tracking policies, loosely after Sorensen (1970) and Gamoran (1992).

Table 4.3: Different tracking measures. Measures based on the age of split: Koerselman (K1, K2); Hanushek and Woessmann (2006) (H1, H2); Waldinger (2006) (W); Pfeffer (Pfeffer2010) (P); Bedard and Cho (2007) (B1), Brunello and Checchi (2007) (C1, C2).

country	K1	K2	H1	H2	W	P	B1	C1	C2
Argentina	0	0	0	0					
Armenia	0	0	0	1					
Australia	0	0	0	0	0		11	16	15
Canada	0	0	0	0	0	0	12	18	0
Cyprus	0	0	0	1					
Denmark	0	0	0	0	0	0	9	16	25
England	0	0	0	0	0	0	11	16	15
France	0	0	0	1	0		9	15	25
Georgia	0	0							
Greece	0	0	0	1	0		9	15	25
Hong Kong	0	0	0	0					
Iceland	0	0	0	0	0		10	16	27
Indonesia	0	0	0	1					
Japan	0	0	0	0	0		9	15	25
Latvia	0	0	0	0				16	25
Macedonia	0	0	0	1					
Moldova	0	0	0	0					
Morocco	0	0	0	0					
New Zealand	0	0	0	0	0	0	11	16	15
Norway	0	0	0	0	0	0	10	16	17
Poland	0	0			0	1		15	38
Portugal	0	0	0	1	0		6	15	25
Romania	0	0	0	1					
Russia	0	0	0	1				15	22
Scotland	0	0	0	0	0	0	11	16	15
Slovenia	0	0	0	1		1		15	33
South Korea	0	0	1	1	0		9	14	33
Spain	0	0			0		10	16	17
Sweden	0	0	0	0	0	0	9	16	25
Taiwan	0	0	0	1					
Thailand	0	0	0	0					
Tunisia	0	0							
Turkey	0	0	0	0	0			11	55

continued on next page

continued from previous page

country	K1	K2	H1	H2	W	P	B1	C1	C2
United States	0	0	0	0	0	0	12	18	0
Yemen	0	0							
Bulgaria	0	1	1	1				14	36
Iran	0	1	0	0					
Italy	0	1	1	1	0	1	8	14	38
Lithuania	0	1	1	1					
Philippines	0	1	1	1					
Singapore	0	1	1	1					
Wallonia	0	1			0	1	7	12	50
Austria	1	1	1	1	1	2	4	10	67
Czech Republic	1	1	1	1	1	1	5	11	62
Flanders	1	1	1	1	0	1	8	12	50
Germany	1	1			1	2	4	10	69
Hungary	1	1	1	1	1	1	4	11	67
Ireland	1	1	1	1	0	0	11	15	81
Israel	1	1	1	1					
Luxembourg	1	1			0			13	46
Netherlands	1	1	1	1	0		8	13	50
Slovak Republic	1	1	1	1	1		4	11	62

Table 4.4: Different tracking measures. Measures based on the percentage of higher track students: Bedard and Cho (2007) (B2), Brunello and Checchi (2007) (C3). Measures based on the number of tracks: Ammermueller (2005) (A).

country	B2	C3	A
Argentina			
Armenia			
Australia	100	63	
Canada	100	7	1
Cyprus			
Denmark	48	53	
England	100	72	3
France	48	56	1
Georgia			
Greece	61	40	2
Hong Kong		43	
Iceland	56	37	1
Indonesia		36	
Japan	75	25	
Latvia			1
Macedonia			
Moldova			
Morocco			
New Zealand	100	37	2
Norway	31	58	1
Poland		61	
Portugal	71	29	
Romania			
Russia		39	2
Scotland	100	72	
Slovenia			
South Korea	58	32	
Spain	66	38	
Sweden	87	50	1
Taiwan			
Thailand		24	
Tunisia		4	
Turkey		39	

continued on next page

continued from previous page

country	B2	C3	A
United States	100	0	
Yemen			
Bulgaria			
Iran			
Italy	33	27	3
Lithuania			
Philippines			
Singapore			
Wallonia	53	70	
Austria	13	72	
Czech Republic	19	80	2
Flanders	38	70	
Germany	26	63	3
Hungary	28	27	2
Ireland	100	24	
Israel		35	
Luxembourg		64	
Netherlands	38	69	
Slovak Republic	24	76	

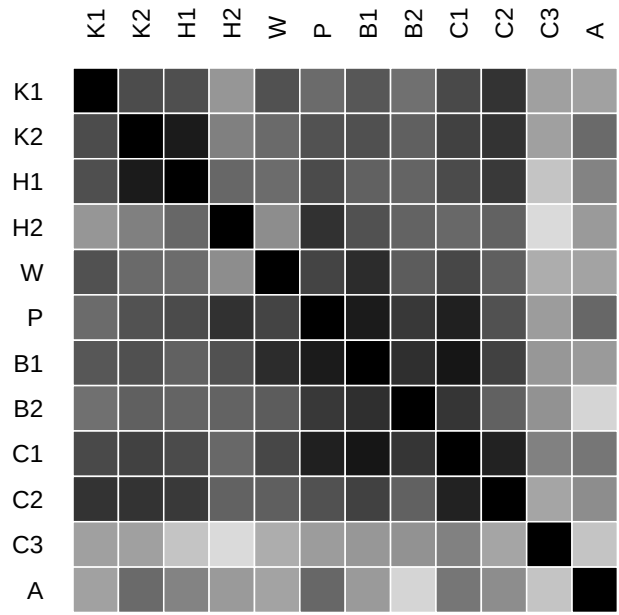

Figure 4.2: Absolute values of correlations between the tracking measures reported in Tables 4.3 and 4.4. A white square indicates a correlation of 0, a black one a correlation of 1 or -1.

4.3 The case for and against tracking

Because of the equalizing nature of peer effects, it should be expected that tracking leads to larger differences in student outcomes. In tracked systems, students attend more homogeneous classes, meaning smaller peer effects and thus larger differences.

Apart from the effects of peers, different tracks will usually also have different curricula. There may also be differences in the number of hours of teaching as well as in other inputs. The separate effects of tracking through peers and other inputs is usually hard to measure. Still, we should in general expect that differences in inputs between tracks are such that they reinforce the differences between them. Exceptions may be remedial classes where some educational inputs are increased to make up for deficiencies in others.

Whether tracking is 'efficient' is a priori unclear. One the one hand, it may be better to teach each student the exact set of skills that he will need in his future job. This argument is perhaps the most pervasive when tracking horizontally. The parallel argument for vertical tracking is to teach to each students at a pace adapted to his ability.

A comprehensive educational system may however still be preferred if it better protects against skill depreciation because of technical change, and this effect outweighs the benefits of specialization (cf. Brunello et al. 2004). On the other hand, skill depreciation can just as well be taken as an argument to widen the upper track (e.g. Maurin and McNally 2007) or to teach a

more academic curriculum in the lower track, rather than to delay tracking.

Both specialization and skill depreciation relate to productivity in a way that is unlikely register on educational tests. Specialized skills may not show up in test scores if we try to measure general skills such as reading and mathematics. Skill depreciation will be much smaller at the time of the test than in the later part of the individuals' working careers.

Another argument in favor of tracking is that resources may be better tailored to the needs of the group in tracked systems. If we believe in the Lazear model on page 48, for example, we could put high track students in larger classes than low track students. In such a world, it may be possible to achieve Pareto improvements.

Theoretical Pareto improvements may however hide the political danger that a narrow, exclusive lower track cannot retain the resources it needs when it no longer draws students from the upper and middle class voters. The tracked outcome may not in fact be an improvement for those attending the lower track, not least because peer effects are relatively large compared to the effects of a redistribution of financial resources. An example of this can be found in Brunello and Checchi (2007), where higher tracks have less students per teacher rather than more.

If the selection mechanism underlying it is good, tracking can prevent 'individuals from raising false hopes towards an unattainable goal.' (Ono 2001) On the other hand, if it is not, tracking may lead to mismatches between students and future jobs. This

is both because of a generally more noisy measurement of individual characteristics (Brunello et al. 2004) and because of the larger impact of parental background when tracking at an earlier age (Bauer and Riphahn 2006; cf. Ammermueller 2005, Waldinger 2006, Jerrim and Micklewright 2010). The latter factor can negatively affect intergenerational mobility (Brunello and Checchi 2007; Maurin and McNally 2008).

In terms of justice, it can be argued that early tracking systems are more fair in a meritocratic sense, especially when track selection is determined by a fair achievement test only. On the other hand, homogeneous comprehensive systems are more just in an egalitarian sense.

Tracked systems with limited track mobility can be seen as particularly unjust to individuals who are misclassified as belonging to the lower track, either because of chance, or because of the effects of parental background.

4.4 Empirical evidence on tracking and efficiency

The European post-war comprehensive school reforms that delayed the age of first tracking have led to considerable debate and research on the effects of tracking. Researchers and policymakers have tried to evaluate the reforms at the time of their implementation, but their efforts have been hindered by a lack of data, especially on later outcomes, and perhaps also by a lack of computing power. This has left room for a new round of analyses in the last decade by researchers using linked information on later outcomes, IV-techniques and modern computers.

Meghir and Palme (2005) and Pekkarinen et al. (2009a, 2009b) look at the comprehensive school reforms in Sweden and Finland respectively. They use variation between subsequent cohorts at the time of a comprehensive school reform. The reform was not implemented everywhere at the same time, and time trends can be controlled for. Meghir and Palme find higher average wages after the reform. Pekkarinen et al. find higher average test scores after the reform, lower test score variation and higher intergenerational mobility.

Unfortunately, comprehensive school reforms tend to include other changes than the postponement of tracking alone, and it is hard to differentiate between their respective effects. For example, it is no surprise that average achievement increases when the lower track is effectively being integrated into the higher one, and the quantity of education is increased for the lower

track students. In the case of Meghir and Palme, an important part of the reform was an increase in the compulsory schooling age, and so the Swedish reform is perhaps better seen as a combined comprehensive schooling and compulsory schooling age reform.

An interesting paper in this respect is that of Ofer Malamud and Christian Pop-Eleches (2007), which looks at a Romanian comprehensive school reform in which the curriculum was changed, but peer group composition was left unaltered. The authors find that although children from disadvantaged backgrounds were more likely to complete the academic track after the reform, they were not more likely to complete university education. Similarly, the Dutch parliamentary report *Parlementair Onderzoek Onderwijsvernieuwingen* (2008) documents how a reform which aimed to implement a common curriculum across a tracked school system, failed. Differences between schools persisted, and the post-reform achievement difference between the tracks was just as large as before.

The UK and South Korea have gone through similar reforms as Sweden and Finland, but unfortunately educational panel data are not available across the reforms, and authors have to rely on cross-sectional snapshots during a time when the reform was implemented in some regions, but not in others.

In the UK, differences between students increased more in pre-reform, tracked regions (see e.g. Kerckhoff 1986, Galindo-Rueda and Vignoles 2004). Galindo-Rueda and Vignoles also find a positive effect of tracking on average scores. Kim et al. (2003)

find a positive effect of tracking on average scores in South Korean data.

International student achievement tests allow for cross-country comparisons. Eric Hanushek and Ludger Woessmann (2006) use international test data collected in PISA, PIRLS and TIMSS since 1995 to estimate the effect of tracking on the distribution of outcomes. They find lower average achievement and higher score variation in early tracking countries.

Within the US, within-school tracking is more relevant. Slavin (1990) reviews a large body of earlier literature, mainly on within-school tracking in the US. He finds that the effect on average scores is close to zero.

There is some experimental evidence on tracking as well. Duflo et al. (2008) use an educational experiment in Kenya to assess the impact of tracking. They find higher average achievement combined with a relatively modest increase in inequality. Even though experimental findings lack many of the identification problems that the studies above have to cope with, Duflo et al. rightly question the generalizability of their results to developed countries.

In summary, most authors seem to agree that tracking increases inequality in educational outcomes (cf. Pfeffer 2009). This is in accordance with our priors. Because of peer effects, tracking should cause larger population-wide differences by reducing within-class heterogeneity, and tracked school systems should display larger inequality in educational achievement than comprehensive school systems.

At the same time, there is considerable variation in the estimated effect of tracking on mean achievement. I will argue further below that these differences can be explained, and are less inconsistent than they may seem at first glance.

authors	effect
comprehensive school reform, panel data	
Pekkarinen et al. (2009a)	–
comprehensive school reform, cross-section	
Kim et al. (2003)	+
Galindo-Rueda and Vignoles (2004)	+
spatial cross-section	
Hanushek and Woessmann (2006)	–
Slavin (1990)	0
experimental	
Duflo et al. (2008)	+

Table 4.5: Important studies on the effect of tracking on mean test scores.

Chapter 5

Incentive effects of curriculum tracking

5.1 Introduction

The literature has mainly focused on the long-term, net effect of curriculum tracking on educational achievement and wages, measuring outcomes after the end of compulsory education or later. I argue that it is also important to look at early-age effects of tracking policies on student outcomes.

Specifically, tracking creates incentives before its start, amongst others for students to work harder in order to get into a higher track. The tracking point is thus a high-stakes moment for the student, whether the track choice is based on an explicit test or

not.

We know from the literature that high-stakes tests should have such effects. Bishop (1998), Jacob (2005) and Neal and Schanzenbach (2010) find that high-stakes tests lead to higher student achievement. This is a subset of a more general literature which shows that students and teachers respond to incentives (Bishop 2006).

The idea that tracking changes incentives is not new. Waldinger (2006) mentions the possible existence of incentive effects. In the model of Eisenkopf (2009), tracking makes educational signaling more efficient by shifting incentives to an earlier age. Galindo-Rueda and Vignoles (2004) find incentive effects in UK data, but the main focus of their paper is on post-tracking outcomes.

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

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

the crowding-out of non-tested ones.

Teachers have an incentive to teach better as well as to substitute time and effort towards tested subjects. It seems a reasonable assumption that teachers should do this for their students' sake, but it may also be in their own interest to do so. The track placement of students (and the possible test preceding it) makes teacher quality more visible, and makes it easier for principals to reward and punish teacher effort as well as easier for parents to choose better schools for their children. Teachers do indeed change their behavior in expected ways in Jacob (2005).

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

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

To at least some degree, incentive effects cause students to do better at tests rather than learn more on an underlying level (cf. Klein et al. 2000, Jacob 2005). This is a problem if we want to use incentives to increase underlying achievement. For

the methodological implications of incentive effects however, the measured scores are more relevant than underlying achievement. Incentive effects can lead to inflated test scores relative to long-term effects of underlying achievement, whether the disparity is caused by temporary bumps in underlying or in measured achievement.

My contribution to the literature is to make a comprehensive empirical investigation into the incentive effects of tracking. The possible endogeneity of early age behavior by forward looking agents is ignored too often in the economics of education literature, and while the idea of incentive effects is sometimes mentioned as an incidental effect of a particular reform, but remains largely unrecognized as a general prinicple. I argue that our prior should be that early age test scores are endogenous with regard to later age policies. By collecting evidence from multiple data sets, I wish to demonstrate that incentive effects are a general effect of tracking.

I use three sources of variation in tracking policies: the large post-war UK and Swedish comprehensive school reforms as well as contemporary cross-country variation in tracking policies. While the UK data allow us to control for unobservables with individual ability, one can condition on municipality and time fixed effects in the Swedish data. Based on this information only, one could conclude that the Swedish estimate will be the most informative, and the cross-country estimate the least.

I will however argue that while the Swedish estimates may or may not accurately reflect the effect of the Swedish comprehen-

sive school reform, the reform involved so many more changes to other policies than tracking, that the UK analysis is the most informative on incentive effects.

5.2 UK evidence for incentive effects

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

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

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

In the end, comprehensive schools were implemented in a region-by-region, school-by-school fashion, both by merging or converting existing schools and by creating new ones. (Government

Circular 10/65, 1965; Benn and Chitty 1996, ch. 1; Kerckhoff et al. 1996, ch. 2)

The survey most appropriate to study the UK reform is the longitudinal National Child Development Study (University of London 2008). 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 them were selected into one of two tracks, while the other part entered the comprehensive school system. I will use the 1958 sweep (at the time called Perinatal Mortality Survey) as well as the 1965, 1969 and 1974 sweeps, when the subjects were 0, 7, 11 and 16 years old.

As can be seen from Table 5.1, 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.

Another 2984 disappear from the sample when we require tracking information to be known. This is mostly due to students attending private schools. A small number of private schools indicate that they are comprehensive in the survey. When I include these as comprehensive, and other private schools as tracked, the empirical results stay virtually unchanged. I therefore only report results excluding private school students.

I also disregard students whose schools turned comprehensive in the very year they took the age 11 test, 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

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

Table 5.1: Number of students in the full NCDS sample, as well as in subsamples with increasingly stringent inclusion conditions. The main reason for missing tracking information is private school attendance.

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

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

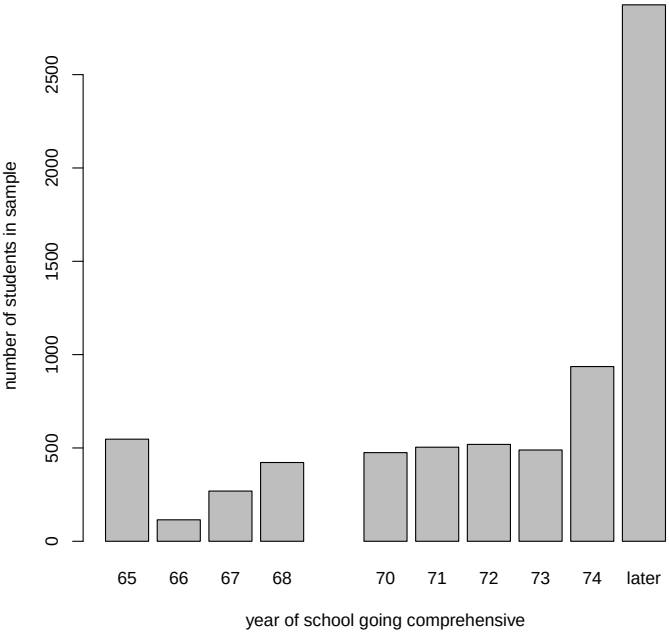

Figure 5.1: Number of students in sample by reform year. The students in the sample all turned 11 in 1969, at which point they were split into tracks in the pre-reform system. Those entering reformed secondary schools (reform year before 1969) should be expected to have lower age 11 scores than those entering schools that reformed after 1969.

dent abilities in different domains.

I synthesize all these variables into one in a two step process. First, I convert all test score distributions to a z-scores because their shapes are arbitrary and skewed, and contain little cardinal level information on underlying achievement. Then, I extract the first principal component of the normalized scores to end up with a measure of general achievement. This process also has the advantage of reducing measurement error from any of the specific tests. Even if the original separate tests yield discrete score distributions, the weighted average of their transformations is smooth enough to be treated as continuous.

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

I encode the tracking status at age 11 (T_s) as a dummy indicating whether the student's school turned comprehensive before 1969, or after. I also select two groups of control variables, listed in Tables 5.4 and 5.5 starting on page 90. The first group A_i consists of standardized age 7 scores and teacher ratings. These include the results of a word recognition and word comprehension test, a copying designs test to assess perceptuo-motor abilities, a draw-a-man test to assess general mental and perceptual abilities, and an arithmetic test.

The second group X_i is a selection of a wide variety of parent and student background variables. I choose not to linearize any of these variables and treat them all as categorical in order to capture as much variation as possible.

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

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

In areas where upper track schools remained, the new comprehensive school may in effect become the new lower track school, with the upper track school attracting all good pupils. Since we can control for ability and background, both forms of selection will lead to an overestimate of incentive effects only to the degree that movers are *unobservably* different in the expected direction.

To take into account the hierarchical nature of the data, I estimate a multilevel or hierarchical linear model (e.g. Gelman and Hill 2007, Pinheiro and Bates 2009) with regressors and error terms on different, nested levels. For the baseline regressions there are two levels: individuals, and LEA×reform year combinations, which I will henceforth call 'schools'. To use LEA×reform year yields slightly larger units, which should lead to more conservative standard errors than using individual schools.

In the first specification

$$y_{s,i} = \alpha + T_s\beta + \varepsilon_s + \varepsilon_i \tag{5.1}$$

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

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

$$y_{s,i} = \alpha + T_s\beta + A_i\gamma + \varepsilon_s + \varepsilon_i \tag{5.2}$$

$$y_{s,i} = \alpha + T_s\beta + A_i\gamma + X_i\delta + \varepsilon_s + \varepsilon_i \tag{5.3}$$

The estimates for these specifications can be seen from the Table 5.2.

The first column shows the unadjusted relationship between age 11 scores and the tracking variable is 0.15 of a UK standard devi-

Dependent variable: UK achievement age 11 (1969)						
specification	(5.1)	(5.2)	(5.3)	(5.4)	(5.5)	(5.6)
tracking (T)	0.15	0.10	0.09	0.10	0.11	0.08
	0.04	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*
ability (A_i)		yes	yes	yes	yes	yes
controls (X_i)			yes	yes	yes	yes
students	7150	7150	7150	5634	7150	7150
grouping	schools	schools	schools	schools	LEAs	years
groups	645	645	645	556	167	10

Table 5.2: Incentive effects in the UK. Students who knew their lower secondary school would be comprehensive score lower than those who had reason to expect a tracked school. Standard errors in italics.

ation. This is a sizable difference, but probably an overestimate of the causal effect.

Turning to column (5.2), we can see that the estimated effect indeed declines to 0.10 when we control for age 7 scores. If we are lucky, the inclusion of age 7 test scores is enough to control for the nonrandom nature of the tracking reforms. In column (5.3), I have added all background variables in X_i as well. The estimate now stands at an only slightly smaller 0.09. This strongly suggests that age 7 test scores pick up most of the selection, and that even less selection is left after the inclusion of X_i.

Even if we can control for the non-randomness of reform areas, we are still left with possible problems of student selection between and within areas. I rerun specification (5.3) to include only students that did not move to a different LEA between ages 7 and 11. This reduces the number of students from 7150 to

5634, and the number of schools from 645 to 556 (the sampling method causes individual schools to be represented by small numbers of students). As can be seen from column (5.4), the estimate even increases a bit to 0.10.

Next, I look at possible selection within areas by using the share of students exposed to a tracked school within each area as the measure of tracking for each student. I define an area as the Local Education Authority: the policy-setting authority of which there are 167 in the sample. As can be seen from column (5.5) however, the point estimate is larger than in the baseline model suggesting that within-LEA selection is not a problem given the controls available to us.

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

$$y_{y,i} = \alpha + T_y\beta + A_i\gamma + X_i\delta + \varepsilon_y + \varepsilon_i \tag{5.6}$$

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

An illustration of this specification can be seen from Figure 5.2. The students on the left side of the figure knew they were going to enter a comprehensive school while those on the right side did not. We can speculate that those attending schools that reformed later had both less uncertainty over the continued tracked status of their secondary school and a larger actual incentive to enter the higher track. Such a pattern is indeed vis-

dependent variable	age 11	age 7
specification	(5.7)	(5.8)
tracking (T)	0.13	0.04
	0.03	*0.04*
controls (X_i)	yes	yes
students	7150	7150
grouping	schools	schools
groups	645	645

Table 5.3: Placebo test for UK incentive effects using early age scores.

ible in the figure. Early test scores are not only larger for those entering a tracked school, but are also increasing in the number of years the reform lies in the future for any particular student.

I also estimate a model with age 7 achievement as the dependent variable as a kind of placebo test under the assumption that incentive effects should be weaker the longer before the tracking point we measure achievement. Since we cannot control for early age scores when using them as the dependent variable, we should expect these estimates to include some selection. Still, as can be seen from Table 5.3, the estimated treatment effect is much smaller and not significantly different from zero for age 7 outcomes. Because I have inflated test scores to account for measurement error, this result is unlikely to be due to the age 7 measurements being more noisy. I interpret the results of the test as additional evidence for the credibility of the original specification.

Do incentive effects differ by gender or background? I add an in-

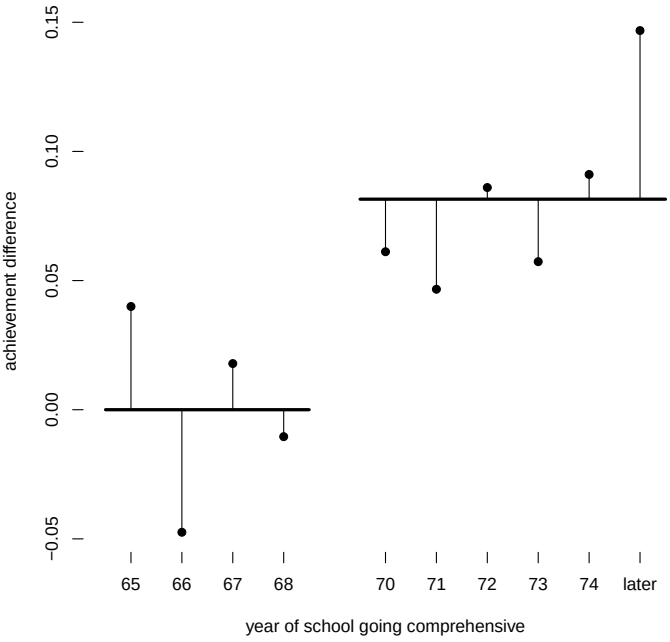

Figure 5.2: Secondary schools left of the divide turned comprehensive before the NCDS students could enter them. Achievement estimates from specification (5.6). Circles indicate the year-level errors.

teraction with gender to specification (5.3). Estimated incentive effects are larger for boys, but not significantly so. I also add interactions on father's socioeconomic status to specification (5.3), but no monotonic pattern can be seen, and the uncertainty of the interactions is large. I have illustrated these results in Figure 5.3.

A quantile regression version of specification (5.3) suggests that incentive effects are slightly larger at the higher end of the distribution. This difference is however not statistically significant. To illustrate this, I have used quantile regression on 100 bootstrap replications of the data, subtracted the estimated effect on the median in every replication, and plotted the 5th, 50th and 95th percentile estimate for every test score quantile in Figure 5.4. This produces an indication of confidence bounds on the slope rather than on the location of the quantile profile.

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

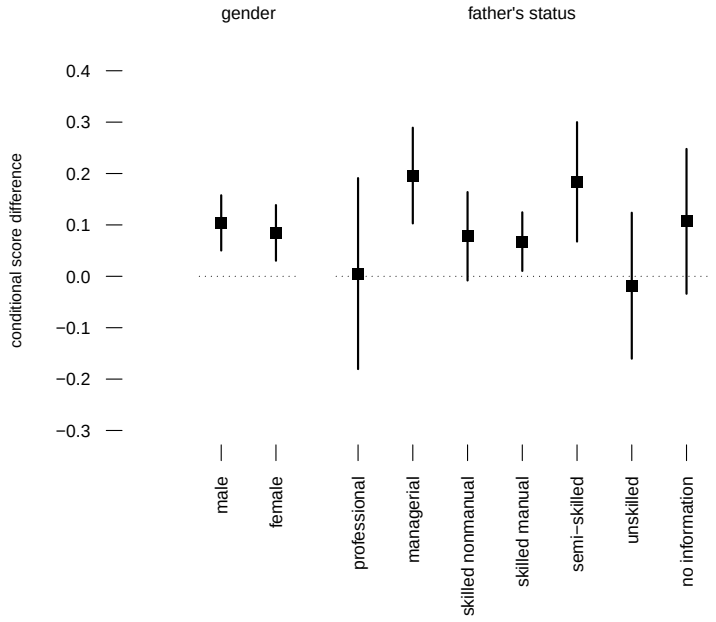

Figure 5.3: Estimated incentive effects for different subgroups. Bars indicate the 95% confidence interval. The size of the effect is not significantly different between boys and girls. No monotonic pattern can be found in the socio-economic background of the student.

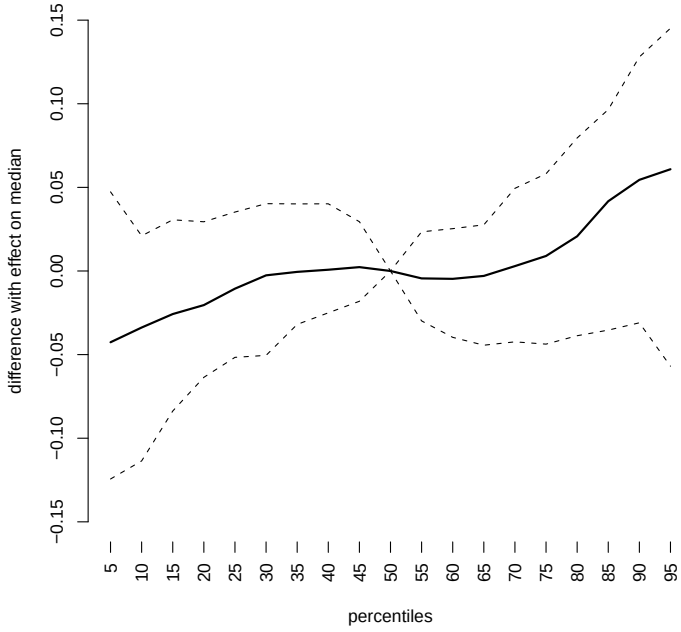

Figure 5.4: Estimated incentive effects relative to the effect at the median for specification (5.3). The dashed lines indicate approximate an 95% confidence interval for the slope of the quantile profile.

	overall		tracked	compr.
variable name	mean	sd	mean	mean
dependent variable y				
Achievement age 11	0.00	1.00	0.04	-0.18
early ability A_i				
Arithmetic score age 7	0.00	1.00	0.01	-0.06
Copying designs score age 7	0.00	1.00	0.01	-0.05
Drawing score age 7	0.00	1.00	0.01	-0.02
Reading score age 7	0.00	1.00	0.03	-0.13
Creativity rating age 7	0.00	1.00	0.01	-0.06
Numbers rating age 7	0.00	1.00	0.03	-0.11
Oral ability rating age 7	0.00	1.00	0.01	-0.05
Reading rating age 7	0.00	1.00	0.03	-0.12
World awareness rating age 7	0.00	1.00	0.02	-0.09

Table 5.4: NCDS student-weighted descriptive statistics of the underlying values of y and A_i.

Table 5.5: NCDS student-weighted descriptive statistics of the underlying values of X_i.

variable name	overall mean	tracked mean	compr. mean
Additional controls X_i			
Female	0.49	0.49	0.50
Height age 11			
1st quintile group	0.19	0.18	0.21
1st quintile group	0.19	0.19	0.17
2nd quintile group	0.19	0.19	0.18
3rd quintile group	0.19	0.19	0.18
4th quintile group	0.19	0.19	0.18
5th quintile group	0.07	0.07	0.08
Father figure			
natural father	0.91	0.91	0.90
other	0.06	0.06	0.07
no information	0.03	0.03	0.04
Father reads to child			
often	0.33	0.34	0.31
occasionally	0.33	0.33	0.34
hardly ever	0.26	0.26	0.27
no information	0.07	0.07	0.08
Mother reads to child			
often	0.46	0.47	0.43
occasionally	0.34	0.33	0.36
hardly ever	0.16	0.15	0.16
no information	0.04	0.04	0.05
Socio-economic status father			
professional	0.04	0.04	0.03
manegerial/technical	0.16	0.17	0.14
skilled nonmanual	0.09	0.09	0.09
skilled manual	0.42	0.42	0.44
semi-skilled	0.16	0.16	0.17
unskilled	0.05	0.05	0.06
no information	0.06	0.06	0.06
Father's education ISCED			
5	0.03	0.03	0.02
3	0.16	0.17	0.15
2	0.54	0.55	0.52

continued on next page

continued from previous page

variable name	overall mean	tracked mean	compr. mean
1	0.01	0.01	0.02
no information	0.25	0.25	0.29
Mother's education ISCED			
5	0.02	0.02	0.01
3	0.19	0.19	0.19
2	0.57	0.57	0.56
1	0.01	0.01	0.01
no information	0.21	0.20	0.23
Father reads books			
often	0.46	0.47	0.42
occasionally	0.20	0.19	0.23
hardly ever	0.27	0.27	0.26
no information	0.07	0.07	0.08
Mother reads books			
often	0.32	0.32	0.29
occasionally	0.21	0.21	0.22
hardly ever	0.42	0.41	0.44
no information	0.05	0.05	0.05
Accomodation type			
house	0.86	0.86	0.84
flat	0.07	0.07	0.08
rooms	0.02	0.01	0.02
no information or other	0.05	0.00	0.00
Father born			
British Isles	0.90	0.90	0.88
Eire or Ulster	0.03	0.03	0.03
other	0.04	0.04	0.05
Mother born			
British Isles	0.91	0.92	0.89
Eire or Ulster	0.03	0.03	0.03
other	0.04	0.03	0.06
Poor at English age 7			
no	0.97	0.98	0.95
somewhat	0.01	0.01	0.02
certainly	0.00	0.00	0.01
no information	0.01	0.01	0.02
Child goes reluctantly to school, age 7			
no	0.86	0.86	0.86

continued on next page

continued from previous page

variable name	overall mean	tracked mean	compr. mean
yes	0.10	0.10	0.10
no information	0.04	0.04	0.04

5.3 Incentives in the Swedish comprehensive school reform

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

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

While there was general agreement that the educational system needed to be improved, the question of whether tracking should be postponed at the same time led to a long 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 dif-

ferentiation 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 both the curriculum, the quality of education and its quantity. In the new system, families also received additional financial support now that they had to keep their children longer in school. (Marklund 1981)

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

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

To see whether any patterns can be found, 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 Swe-

den (see Harnqvist 2000). 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.

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

I have data on spatial, verbal and inductive components of an age 12 ability test for most students, as well as standardized tests in mathematics for those who were in 6th grade of primary school. I transform each 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 have at least some information for 21877 students in 1020 municipalities in the full sample. As can be seen from Table 5.6, this decreases to 19946 students in 1013 municipalities for which I have information on IQ, and further to 17427 students in 1005 municipalities for those which I have math scores as well.

While it may not be all too far from the truth that the students without IQ scores were missing at random, the students with IQ scores but without a mathematics test score are not a random

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

Table 5.6: Number of observations in the full UGU 1948 and 1953 sample, as well as in the subsample with known ability scores and ability and mathematics scores respectively. As can be seen from the last four rows, the panel of municipalities is not balanced.

selection. They partly consist of those that either were not in the 6th grade when their peers were, and of those that had transferred to an upper track school at an earlier age. I will look at the effects of excluding this group further below.

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

I consider two families of models. In the fixed effects models

$$y_i = \alpha + T_i\beta + MC_i\gamma + X_i\delta + Z_i\zeta + \varepsilon_i \tag{5.10}$$

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

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

Table 5.7: Estimates of the effects of the Swedish comprehensive school reform on early test scores.

standard errors clustered on the municipality×cohort level.

An alternative family of models uses county and cohort fixed effects, with separate intercepts for the three largest cities, and municipality×cohort random effects.

$$y_{mc,i} = \alpha + T_{mc}\beta + CC_{mc}\gamma + X_{mc}\delta + Z_i\zeta + \varepsilon_{mc} + \varepsilon_i \quad (5.11)$$

The random effects model can be more efficient, but it cannot control for potential municipality level selection.

To test for bias in the random effects models, I do a Hausman specification test under the assumption that the fixed effects model is consistent. For most specifications I reject the null that the random effects model is consistent at the 5% level. I therefore only report results from the fixed effects models below.

I have listed estimation results in Table 5.7. As can be seen from the first column, there seems to be a significantly negative conditional relationship between tracking and IQ, while we can see from the second column that the coefficient on math scores

is not significantly different from zero.

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

One explanation could be that measurement error is much larger for mathematics than for IQ. Unfortunately, there is not enough information in the UGU data set to check for this.

Another possibility is that the sample is not representative of the student population in each municipality. Mathematics scores are only known for those students who were in the 6th grade of either the new comprehensive school or of the old primary school. Missing mathematics scores have two effects on the estimates.

I rerun the IQ regression of the second column on a sample including the students with missing mathematics scores. As can be seen from the fourth column of Table 5.7, selectively missing students within municipalities seem to be able to explain most of the negative conditional correlation between tracking and IQ.

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

Under the assumption that the true effect of tracking on IQ is

zero, we can use it to control for selection. As can be seen from the third column in the table, the estimated effect of the reform on math scores conditional on ability scores is very close to zero.

The best estimate of the reform effect on IQ comes from the fourth column in Table 5.7, and the best estimate of the effect on mathematics skills comes either from the second or the third, depending on assumptions. None of these three estimates is significantly different from zero. It is possible to obtain borderline significantly positive or negative effects with other model variations, but these results are never robust to small and arbitrary model changes.

It is possible that the lack of clear results are due to measurement error in the reform variable. To check on this, I merge the UGU data with reform years by municipality which Holmlund (2007) has collected. I obtain point estimates close to the estimates in Table 5.7, and I conclude that measurement error is not likely to be the main driver of these results.

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

policy experiment than the Swedish one.

5.4 International evidence for incentive effects

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

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

I take tracking information mainly from the Eurybase database (Eurydice 2008), supplemented with information from Wikipedia

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

As control variables, I use real per capita purchasing power-adjusted GDP (expressed in 10 000 USD) from the Penn World Table (2006) as well as educational expenditures as a percentage of GDP from the World Bank EdStat database (2011). For GDP, the year of observation is always 1995. For educational expenditures, it is the available observation the closest to 1995. Descriptive statistics for these and other variables can be seen from Table 5.8. I have complete data on 1040596 students in 51 countries.

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

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

Table 5.8: International data: descriptive statistics for the full sample (top), and for the EEA countries only (bottom).

This reduces the sample to 515788 students in 28 countries.

As in Section 5.2, I estimate a multilevel model to take into account the errors individuals have in common when they share a class, school or country. The error structure in all specifications is nested, and given by

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

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

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

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + \varepsilon \tag{5.12}$$

The results from estimating this equation can be seen from column (5.12) in Table 5.9. Countries with early tracking clearly have higher score means, with the mean difference as large as 0.41 standard deviations of international student test scores.

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

Dependent variable: international early age achievement					
	(5.12)	(5.13)	(5.14)	(5.15)	(5.16)
tracking (T)	0.41	0.17	0.23	0.23	0.26
	0.19	*0.16*	*0.07*	*0.07*	*0.07*
GDP		0.38	-0.01	-0.02	-0.01
		0.08	*0.05*	*0.05*	*0.05*
expenditures		-0.08	0.03	0.02	0.03
		0.04	*0.03*	*0.02*	*0.03*
books at home				0.14	
				0.00	
$T\times$books at home				0.00	
				0.04	
female					0.04
					0.00
$T\times$female					-0.05
					0.02
students	1040596	1040596	515788	515788	515788
countries	51	51	28	28	28

Table 5.9: International evidence for incentive effects; pooled multilevel regression based on international data. Standard errors in italics.

level matrix C_{cn}.

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + C_{cn}\delta + \varepsilon \tag{5.13}$$

The estimates from this specification can be seen from column (5.13). Estimated incentive effects are now more than halved at 0.17 standard deviations.

However when we turn to the EEA sample, the estimate is improved in many ways. It can be seen from column (5.14). GDP and educational expenditures now play a much smaller role, both turning statistically insignificant.

At 0.23, the estimated incentive effects are now larger, but also

much more precisely estimated. This is exactly what we should expect if the tracking variable has classical measurement error for non-EEA countries. Another indication that this is the better estimate is that the estimated effect of educational expenditures now has the expected sign, even if it is insignificant.

I have illustrated the estimate from specification (5.14) in Figure 5.5. As can be seen from the figure, a specification linear in age may seem to fit the data better, but the results would become more sensitive to the exact tracking ages we assign to late tracking countries.

The estimate is still not likely to reflect a causal effect in the sense that a country that randomly decides to change its tracking policies is likely not to experience a change in early test scores as large as 0.23 international standard deviations. One can easily imagine that the pattern is a combination of incentive effects of tracking, and a tendency for countries that stress the importance of achievement on hard, testable subjects in primary school to have retained a tracked secondary school system. The remarkably strong pattern in Figure 5.5 does however suggest that early tracking and early achievement are strongly related.

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

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + C_{cn}\delta + B_i\theta + (B_i \cdot T_{cn})\kappa + \varepsilon \quad (5.15)$$

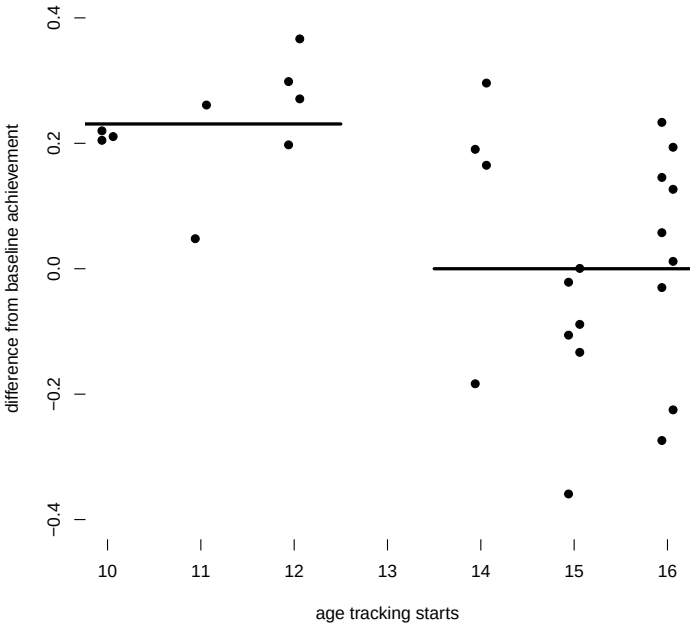

Figure 5.5: An illustration of the EEA estimate of incentive effects from specification (5.14). Early tracking countries have higher conditional early test scores. The solid line represents the estimate, dots indicate the country-level errors. The horizontal axis has been jittered slightly to improve visibility.

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

Results can be seen from column (5.15). Students with more than one case of books at home score higher on average, but the interaction with tracking is insignificant and close to zero.

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

$$y_{cn,s,cl,i} = \alpha + T_{cn}\beta + D_s\gamma + C_{cn}\delta + F_i\lambda + (F_i \cdot T_{cn})\mu + \varepsilon \quad (5.16)$$

Looking at column (5.16) of Table 5.9, we can see that the differences between boys and girls are moderately small at -0.05. Both the unclear differences in parental background and the smaller point estimate of incentive effects for girls mirror the UK findings.

Hanushek and Woessmann make a slightly different assessment of the tracking age, even if they define tracking in the same way. A re-run of my regressions with an age 14 tracking dummy based on the Hanushek and Woessmann variable gives higher and more precise point estimates in specifications (5.12) and (5.13), but makes no difference in the EEA sample of the later specifications.

All in all, international test score data provide us with an additional line of evidence for incentive effects. The estimated effect is unlikely to reflect an unidirectional causal link between track-

ing and early test scores only, but the relationship is nevertheless exceptionally clear.

5.5 Discussion

In this chapter, I have looked for incentive effects of curriculum tracking in the UK, across countries and in Sweden. 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.

I find empirical evidence to support this hypothesis. In UK data, tracking seems to cause an incentive effect of 0.09 UK standard deviations. Within the European Economic Area, tracking is associated with 0.23 international standard deviations higher scores. In the British case the estimates are likely to reflect a causal mechanism from the comprehensive school reform on test scores, while the international estimates capture both the effects of tracking on achievement and the effects of culture on both. These estimates are large, but not larger than the 0.2–0.3 Jacob (2005) finds for a high-stakes test.

The Swedish results are much more unclear. We are probably seeing the effects of both multiple causal mechanisms and multiply types of selection at the same time. It is hard to draw any conclusions on incentive effects. Both positive and negative incentive effects are consistent with the data under different assumptions.

Incentive effects have methodological implications. The existence of incentive effects makes value-added estimates of the later age effects of tracking (e.g. Hanushek and Woessmann

2006, cf. Todd and Wolpin 2003) misspecified. Pre-tracking test scores are not exogenous, but positively related to early tracking, leading to a downward bias in tracking estimates that use early test scores to control for unobservables.

I have illustrated this in Figure 5.6. If we compare an early tracking to a late tracking country, the late tracking system may appear to be more successful because it seems to catch up to the early tracking country during the years in which tracking policies differ. This is however not the relevant comparison because the reason why the late tracking country had lower early test scores in the first place is exactly because it tracks late.

If we accept the invalidity of value-added specifications, we can reconcile previous studies on the long-term effect of tracking. I have listed some current papers on the mean effect of tracking in Table 5.10. The effect on the mean is negative in Pekkarinen et al. as well as in Hanushek and Woessmann.

We should not be surprised to find an apparent negative effect of tracking in studies of postwar reforms such as Pekkarinen et al. The reforms simultaneously changed the tracking structure and improved the quantity and quality of education of those previously in the lower track, and we are measuring the effects of both. A policy experiment more relevant to tracking policies today would be if a country with a modern vocational track such as Germany were to postpone its tracking point. The positive effects of such a reform on mean test scores could be much smaller.

The other main study finding a negative effect is that of Hanu-

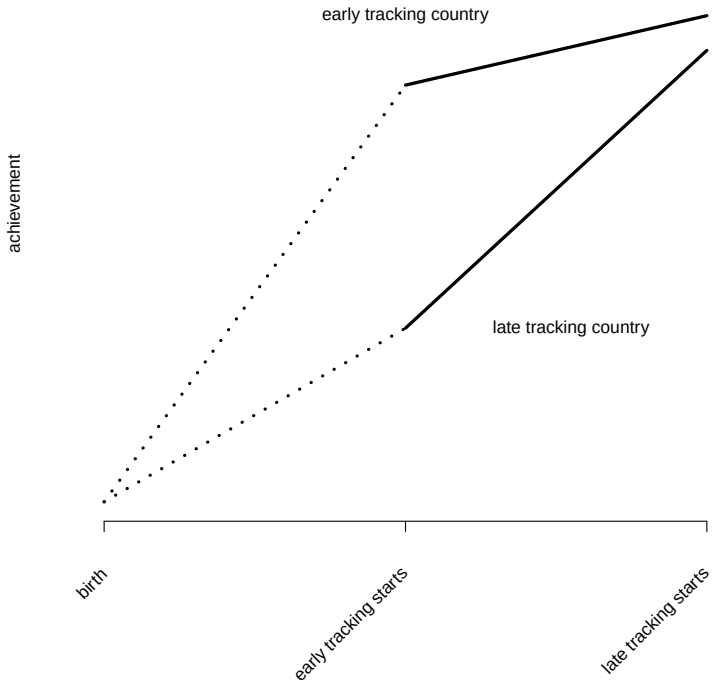

Figure 5.6: Bias in value-added estimates of the later age effects of tracking.

authors	effect
comprehensive school reform, panel data	
Pekkarinen et al. (2009a)	–
comprehensive school reform, cross-section	
Kim et al. (2003)	+
Galindo-Rueda and Vignoles (2004)	+
spatial cross-section	
Hanushek and Woessmann (2006)	–
Slavin (1990)	0
experimental	
Duflo et al. (2008)	+

Table 5.10: Important studies on the effect of tracking on mean test scores.

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

The other authors all find a zero or positive effect of tracking on mean scores. I thus conclude that a zero or positive effect of tracking on mean test scores is the most consistent with the data.

A second implication is that a positive relationship between pre-tracking scores and tracking policies cannot be used as an argument that there is selection in post-tracking regressions (e.g. Manning and Pischke 2006). Since there is no reason to assume that early test scores are exogenous, this kind of placebo test is

not informative of selection problems.

Tracking should be expected to change the level of competitiveness in early education. This can make tracking relevant as a policy tool to regulate the level of incentives. We may feel that incentives are too small in some cases, for example if the labor market returns to effort are much lower than societal returns. This can be the case if income taxes are strongly progressive. In such cases we may want to start tracking at an earlier age in order to increase student competitiveness.

At other times, we may feel that incentives are too strong at too young an age for student well-being. In such cases we may want to postpone tracking in order to lower the burden on young students.

Of course, tracking does not only change incentives, but also actual learning dynamics during the tracked years and possibly beyond. If we are considering to track at an earlier age than is done today in any given country, we may want to combine such a policy change with other measures, for example to increase track mobility (cf. section 4.2).

It would be interesting to look at how persistent early age incentive effects are throughout the student career. Incentive effects are however hard to disentangle empirically from the effects of the reforms themselves. What would be needed is a data set spanning a reform in which some students expected to be tracked, but were not. I am not currently aware of such data.

Chapter 6

Concluding remarks

Comprehensive schooling has an equalizing effect on educational outcomes. Because the associated loss of efficiency seems relatively small and uncertain, there are strong egalitarian arguments in favor of comprehensive middle schools. On the other hand, there are classical liberal and meritocratic arguments in favor of tracking. Students should be allowed to apply to the schools they want to, and schools should be allowed to select among their applicants in a transparent and meritocratic way.

The choice between egalitarianism and meritocracy can be a false one when it comes to educational segregation. Because the incentive for good students to attend the same school is so large, egalitarian late-tracking policies should be expected to fail except in heavily regulated or sparsely populated places.

In a second-best world where egalitarian policies are politically

impossible, we may instead want to combine a transparent and meritocratic early tracking system with other policies that increase intergenerational mobility and reduce inequality.

When considering tracking policies, it should be remembered that tracking is a multidimensional phenomenon. The advantages and disadvantages of tracking may not be proportionately distributed over all dimensions, and different combinations of tracking policies can have non-additive effects. It could for example be the case that we would like to have horizontal tracking, but not vertical, or that we would like to have early vertical tracking in combination with a mechanism that facilitates easy transitions between tracks.

In a similar fashion, incentives may be changed in other ways than through tracking. In fact, the prominence of international educational surveys such as PISA and TIMSS in the public debate may lead to an international shift towards educational policies that stresses individual, class and school level scores on testable subjects. This makes it even more important for both policy makers and policy evaluators to understand what such scores can and can not tell us.

Svensk sammanfattning

Centrala teman i denna avhandling är parallellskolesystem, linjedelning och segregering i högstadier och gymnasier. Det finns stora skillnader mellan länder gällande den årskurs då eleverna sorteras in i olika nivåer genom skilda klasser eller skolor. I Tyskland och Nederländerna till exempel delas eleverna upp före högstadiet medan Sverige och Finland har enhetsskolor på högstadienivå.

Skillnaderna finns inte bara mellan länder, men också inom länder över tid. Länderna som har enhetsskolor idag har ofta infört dem i en serie av grundskolereformer på 50-, 60- och 70-talet. Dessa reformer gör det lättare att studera parallellskolesystemets effekter.

I denna avhandling argumenterar jag att parallellskolesystem har en positiv effekt på elevernas testresultat i årskurserna före linjedelningen. Eleverna har incitament att jobba hårdare för

att komma på i den akademiska linjen (den tidigare realskolan). Jag undersöker incitamentseffekter empiriskt, och hittar mönster konsistenta med incitamentseffekter både i brittisk och i internationell data.

Det andra bidraget i avhandlingen är metodologiskt. Nationalekonomer brukar behandla testresultat från t.ex. IQ-tester eller internationella PISA-undersökningar som kardinala. Testresultat är dock fundamentalt ordinala, och kardinala statistiska operationer såsom nationella genomsnitt behöver varken vara meningsfulla eller robusta.

Jag visar att normalfördelade testresultat ligger tillräckligt nära dess pengavärde på arbetsmarknaden för att kunna tolkas som kardinala. I vissa fall är fördelningarna av testresultat dock mycket sneda, och då kan det vara bra att transformera testresultaten före man använder dem.

Bibliography

D. Acemoglu and J. Angrist. How large are human-capital externalities? Evidence from compulsory schooling laws. *NBER macroeconomics annual*, 15:9–59, 2000.

J. Altonji and C. Pierret. Employer Learning and Statistical Discrimination. *Quarterly Journal of Economics*, 116(1):313–350, 2001.

A. Ammermueller. Educational opportunities and the role of institutions. ZEW discussion paper no. 05-44, 2005.

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

J. Angrist and A. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

O. Ashenfelter, H. Colm, and H. Oosterbeek. A review of es-

timates of the schooling/earnings relationship, with tests for publication bias. *Labour Economics*, 6:453–470, 1999.

P. Bauer and R. Riphahn. Timing of school tracking as a determinant of intergenerational transmission of education. *Economic Letters*, 91:90–97, 2006.

G. Becker. *Human capital: A theoretical and empirical analysis, with special reference to education.* University of Chicago Press, 1964/1993.

K. Bedard and I. Cho. The gender test score gap across OECD countries. Working paper, September 2007.

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

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

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

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

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

G. Brunello, M. Giannini, and K. Ariga. The optimal timing of school tracking. IZA discussion paper no. 995, 2004.

P. Carneiro, J. Heckman, and E. Vytlacil. Understanding what instrumental variables estimate: Estimating marginal and average returns to education. Working paper, 2005.

P. Carneiro, J. Heckman, and E. Vytlacil. Estimating marginal returns to education. NBER working paper no. 16474, 2010.

R. Chakrabarti. Do vouchers lead to sorting under random private school selection? Evidence from the Milwaukee voucher program. *Harvard University*, 2005.

F. Cunha, J. Heckman, and S. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

M. Davison and A. Sharma. Parametric statistics and levels of measurement. *Psychological Bulletin*, 104(1):137–144, 1988.

E. Duflo, P. Dupas, and M. Kremer. Peer effects and the impact of tracking: Evidence from a randomized evaluation in kenya. NBER Working Paper no. 14475, 2008.

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

D. Epple and R. Romano. Competition between private and public schools, vouchers and peer-group effects. *The American Economic Review*, 88(1):33–62, 1998.

D. Epple, E. Newlon, and R. Romano. Ability tracking, school competition and the distribution of educational benefits. *Journal of Public Economics*, 83(1):1–48, 2002.

D. Epple, D. Figlio, and R. Romano. Competition between private and public schools: testing stratification and pricing predictions. *Journal of Public Economics*, 88(7-8):1215–1245, 2004.

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

D. Figlio and M. Page. School choice and the distributional effects of ability tracking: does separation increase equality? NBER working paper no. 8055, 2000.

M. Friedman. *Capitalism and freedom.* Chicago, 1962.

F. Galindo-Rueda. Employer Learning and Schooling-Related Statistical Discrimination in Britain. IZA DP 778, 2003.

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

A. Gamoran. The variable effects of high school tracking. *American Sociological Review*, 57(6):812–828, 1992.

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

Z. Griliches. Economic data issues. *Handbook of econometrics*, 3:1465–1514, 1986.

D. Hand. *Measurement theory and practice*. Oxford University Press, 2004.

H. Hansmann. Higher education as an associative good. *Yale Law and Economics working paper*, 231, 1999.

E. Hanushek. School resources. *Handbook of the Economics of Education*, 2:865–908, 2006.

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

E. Hanushek and L. Woessmann. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. NBER working paper no. 14633, 2009.

E. Hanushek and L. Zhang. Quality-consistent estimates of international schooling and skill gradients. *Journal of Human Capital*, 3(2):107–143, 2009.

J. Hartog. On human capital and individual capabilities. *Review of Income and Wealth*, 47(4):515–540, 2001.

J. Hartog and H. Van den Brink. *Human capital: advances in theory and evidence*. Cambridge University Press, 2007.

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

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

C. Hoxby and G. Weingarth. Taking race out of the equation: School reassignment and the structure of peer effects. *Unpublished manuscript*, 2005.

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

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

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

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

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

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

J. Jerrim and J. Micklewright. Children's cognitive ability and the socioeconomic gradient: changes over age in nine countries. Working paper, 2010.

A. Kerckhoff. Effects of ability grouping in British secondary schools. *American Sociological Review*, 51(6):842–858, 1986.

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

B. Kiker. *Human capital: in retrospect*. Essays in Economics 16, University of South Carolina, 1968.

T. Kim, J. Lee, and Y. Lee. Mixing versus sorting in schooling: evidence from the equalization policy in South Korea. KDI School Working Paper no. 03-07, 2003.

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

A. Krueger and M. Lindahl. Education for growth: why and for whom? *Journal of Economic Literature*, 39(4):1101–1136, 2001.

F. Lange. The speed of employer learning. *Journal of Labor Economics*, 25(1):1–35, 2007.

V. Lavy and A. Schlosser. Mechanisms and impacts of gender peer effects at school. Working paper, May 2007.

V. Lavy, O. Silva, and F. Weinhardt. The good, the bad and the average: Evidence on the scale and nature of ability peer effects in schools. *NBER working paper no. 15600*, 2009.

E. Lazear. Educational production. *The Quarterly Journal of Economics*, 116(3):pp. 777–803, 2001.

E. Lazear. Teacher incentives. *Swedish Economic Policy Review*, 10(2):179–214, 2003.

F. Lord. On the statistical treatment of football numbers. *American Psychologist*, 8:750–751, 1953.

Frederic Lord. *Applications of Item Response Theory to Practical Testing Problems*. L. Erlbaum, 1980.

O. Malamud and C. Pop-Eleches. The effect of postponing tracking on access to higher education: evidence from a regression-discontinuity design. Working paper, October 2007.

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

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

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

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

E. Maurin and S. McNally. Educational effects of widening access to the academic track: a natural experiment. IZA discussion paper no. 2596, 2007.

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

C. Meghir and M. Palme. Educational reform, ability and family background. *American Economic Review*, 95(1):414–424, 2005.

J. Mincer. Schooling, experience, and earnings. *NBER Books*, 1974.

E. Moretti. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of econometrics*, 121(1-2):175–212, 2004.

R. Murnane, J. Willett, Y. Duhaldeborde, and J. Tyler. How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, 19(4):547–568, 2000.

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

Organization for Economic Co-operation and Development OECD. Programme for International Student Assessment PISA. 2006.

Organization for Economic Co-operation and Development OECD. Education at a Glance 2010. 2010a.

Organization for Economic Co-operation and Development OECD. OECD Factbook 2010. 2010b.

Commissie Parlementair Onderzoek Onderwijsvernieuwingen. *Parlementair Onderzoek Onderwijsvernieuwingen*. Sdu Uitgevers, 2008.

H. Ono. Who goes to college? Features of institutional tracking in Japanese higher education. *American Journal of Education*, 109(2):161–195, 2001.

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

T. Pekkarinen, R. Uusitalo, and S. Kerr. School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics*, 93(7-8):965–973, 2009b.

F. Pfeffer. Equality and quality in education. presented at the Youth Inequalities Conference, University College Dublin, Ireland, December 2009.

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

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

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

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

T. Schultz. Investment in human capital. *The American Economic Review*, 51(1):1–17, 1961.

R. Slavin. Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of educational research*, 60(3):471, 1990.

A. Sorensen. Organizational differentiation of students and educational opportunity. *Sociology of Education*, 43(4):355–376, 1970.

R. Speakman and F. Welch. Using wages to infer school quality. *Handbook of the Economics of Education*, 2:813–864, 2006.

M. Spence. Job market signaling. *Quarterly Journal of Economics*, 87(3):355 – 374, 1973.

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

S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

J. Stiglitz. The theory of "screening," education, and the distribution of income. *The American Economic Review*, pages 283–300, 1975.

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

R. Topel. Labor markets and economic growth. *Handbook of labor economics*, 3:2943–2984, 1999.

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

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

U.S. Bureau of Labor Statistics. National Longitudinal Study of Youth 79. 2010.

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

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

World Bank. EdStat Education Statistics. 2011.

A. Zand Scholten and D. Borsboom. A reanalysis of Lord's statistical treatment of football numbers. *Journal of Mathematical Psychology*, 53:69–75, 2009.

Åbo Akademi University Press

ISBN 978-951-765-620-7
ISBN 978-951-765-621-4 (digital)