

School tracking and development of cognitive skills¹

Sari Pekkala Kerr (Wellesley College)²

Tuomas Pekkarinen (Aalto University, IZA and IFAU)

Roope Uusitalo (Government Institute for Economic Research, IZA and IFAU)

[This version Nov 10, 2010]

Abstract

We evaluate the effects of selective vs. comprehensive school systems on mathematical, verbal and logical reasoning skills using data from the Finnish comprehensive school reform. This reform replaced the old two-track school system with a uniform nine-year comprehensive school. We use a differences-in-differences approach and exploit the fact that the reform was implemented gradually across the country during a six-year period. As a measure of cognitive skills we use the test scores in the Finnish Army Basic Skills Test, which is taken at the beginning of the mandatory armed service. We find that the reform had a small positive effect on the verbal test scores, but no effect on the mean performance in the arithmetic or logical reasoning tests. However, the reform significantly improved the scores of the students whose parents had only basic education.

JEL Codes: H52, I21

Key Words: Education, school system, tracking, comprehensive school, test scores

¹ The authors would like to thank Lidia Farré, Sandra McNally, Eva Mörk, William Kerr, David Autor, as well as seminar participants at London School of Economics, Universitat de Alicante, ESPE Conference in London, EALE Conference in Amsterdam for helpful comments. Pekkarinen is grateful for financial assistance from the Academy of Finland.

² Corresponding author. Address: Wellesley College, WCW, 106 Central Street, Wellesley, MA 02481. Email: skerr3@wellesley.edu.

1. Introduction

International comparisons of student achievement, such as the OECD's Programme for International Student Assessment (PISA), have generated a growing interest in the effect of school systems on student outcomes. According to these comparisons, the differences in average test results across countries with roughly equal school resources are very large. Also the dispersion of test scores varies considerably across countries.

One potential explanation for cross-country differences in the level and, in particular, in the variance of achievement scores relates to the extent and timing of tracking or ability grouping of students. For example, the OECD has repeatedly argued that the average variation in student performance tends to be higher in countries with early tracking policies (OECD 2003). High variance of student achievement and its correlation with family background has been seen as problematic from the point of view equality of opportunity. On the other hand, many commentators have raised the concern that postponing tracking too late could have negative effects on test scores by lowering the quality of teaching. Implicit in this debate is an efficiency-equity trade-off: postponement of tracking should improve the equality of opportunity with the potential cost of deterioration of the achievement for some students.

Comprehensive and selective school systems differ importantly when it comes to the timing of tracking. In the selective system, tracking students into different types of schools occurs early and choices made as early as age ten largely determine later schooling options. In the comprehensive system, students often stay in the same schools until the end of secondary school. Hence, the debate on school tracking is inherently entwined with the school system debate.

This study contributes to this debate by estimating the effect of the Finnish comprehensive school reform on cognitive skills tests that are taken at the beginning of the mandatory military service by all male conscripts. Finland had a selective two-track school system until 1970's when the school reform replaced the old two-track system with a uniform comprehensive school that is similar to comprehensive school systems in other European countries. As a result of the reform the tracking age was postponed from age 10 to 15. The differences between the pre- and post-reform systems are similar to the cross-country differences in school systems in the OECD countries today. The effects of the Finnish reform are therefore informative for the current schooling policy debate.

Previous studies such as Meghir and Palme (2005) as well as Pekkarinen et al. (2009) have shown that comprehensive school reforms did improve the equality of opportunity by decreasing the intergenerational correlation of earnings. However, these studies have not addressed the effect of comprehensive school on the distribution of learning outcomes which is a key element in understanding the potential mechanisms by which the school system affects adult outcomes. The earnings effects reported in previous studies could be due to peer effects, social networks, opening of new educational opportunities or direct impact on productive skills. Using a previously unavailable unique test score data from the Finnish Defense Forces we can partially open the ‘black box’ relating school systems to labor market outcomes. This is particularly important given that recent research, surveyed by Hanushek and Wößmann (2009), emphasizes that individual earnings, the distribution of income, and economic growth are most strongly correlated with the distribution of cognitive skills in the population and not just the level of educational attainment.

Economic theory provides somewhat ambiguous predictions on the effect of the comprehensive school system on student achievement. On one hand, a comprehensive system increases heterogeneity in the classroom and probably makes classes more difficult to teach, thereby lowering student achievement (Lazear, 2001). If peer-effects are important, however, any changes in the class composition may also affect student achievement. The effect on the mean achievement depends on whether good students are harmed by bad students more than bad students benefit from being around good students. However, even if the effect on the average student achievement is ambiguous, a comprehensive school system should decrease the variance of test scores, if average peer quality has a positive effect on student outcomes. Furthermore, as early educational choices are more likely to be determined by family background, (Brunello and Checchi, 2007) it seems reasonable to expect the postponement of tracking in the comprehensive system to reduce the correlation of test scores and family background.

Most convincing empirical evidence on the effects of ability tracking on test scores comes from a randomized experiment in Kenya where randomly selected schools implemented tracking and non-tracking policies. Duflo et al. (2008) show that tracking within schools seems to benefit all students. However, it is not clear whether these results can be generalized to developed countries where the student population is typically less heterogeneous. Furthermore, in selective systems students are typically not tracked within schools but to

different types of schools, which necessarily implies that teacher quality and curriculum may vary considerably across the tracks. Debate on the relative advantages of different school systems is primarily concerned about tracking into different types of schools rather than tracking within schools. Hence, even a well designed randomized experiment of tracking within schools is unlikely to settle the policy question of whether the entire school system should be selective or comprehensive.

In developed countries, most of the existing evidence on the potential benefits of selective versus comprehensive system originates from cross-country comparisons. For example, Hanushek and Wößmann (2006) use data from international student assessments and find that the variance in test scores is higher in countries where tracking takes place at an early age. At the same time, early tracking seems to have generally negative effects on mean performance, though the evidence on these mean effects is less consistent. A similar cross-country approach is used by Brunello and Checchi (2007) and Waldinger (2006). Neither of these studies confirms the Hanushek–Wößmann result.

These conflicting results from previous studies reflect, in part, the difficulties in analysing the effects of school system using cross-country data. While these studies try to control for variation due to other factors by including early test scores (Hanushek and Wößmann, 2006; and Waldinger, 2006) or by using time variation in the tracking age (Brunello and Checchi, 2007), it is far from clear that all relevant cross-country differences could reliably be accounted for.

Analyzing the changes in test scores when a country switches from a tracked to a comprehensive system appears to be a more promising approach to identify the effects of the school system on student achievement. Previous attempts to do this include Kerckhoff et al. (1996) as well as Galindo-Rueda and Vignoles (2005), both of whom study the effect of a gradual movement from a selective school system to a comprehensive system in England. Both studies examine the effects of comprehensive system on cognitive ability using test scores from the National Child Development Survey (NCDS). However, as noted by Manning and Pischke (2006), the areas that first switched to the comprehensive system in England were on average poorer than the areas which retained the tracked system. It is therefore difficult to distinguish the effect of school systems from the regional differences using single cohort data such as the NCDS. Manning and Pischke also demonstrate that even value-added

models that control for earlier test scores are likely to be biased as long as regional differences cannot be fully controlled for.

Relative to earlier studies, the distinct advantage of the Finnish reform is the availability of panel data from several cohorts, which avoids the need to rely on the cross-sectional variation only. The Finnish comprehensive school reform was implemented gradually region by region between 1972 and 1977. This gradual implementation allows controlling for regional variation and any time trends in student achievement using a difference-in-differences approach, which avoids biases such as those discussed by Manning and Pischke. Furthermore, the data also include information on families, which makes it possible to estimate the effect of the reform using data on brothers who were placed into different school systems because they differ in age.

We use the same identification strategy as in Pekkarinen et al. (2009) but evaluate the effects of a school reform directly on the distribution of skills that the students are supposed to learn in school. The reform had a small positive effect on the verbal test scores, but little effect on the mean performance in the arithmetic or logical reasoning tests. However, the reform significantly improved the scores of the students whose parents had only basic education

The rest of the paper proceeds as follows. In the next section, we describe the content and the implementation of the Finnish comprehensive reform. The third section describes the data and the Finnish Army Basic Skills Test that we use as a dependent variable. We then move on to present the differences-in-differences and maximum likelihood estimation of the effect of the reform on test scores and in the fifth section we discuss the results. The sixth section concludes.

2. Comprehensive school reform³

2.1 Background

Finland introduced a wide-ranging comprehensive school reform in the 1970's. Similar reforms had already taken place in Sweden in 1950s and in Norway in 1960s (Meghir and Palme, 2005; Aalvik et al. 2003). The Finnish comprehensive school reform abolished the old two-track school system and created a uniform nine-year comprehensive school. The main

³ This section draws on Pekkarinen et al. (2009).

motivation of the reform was to provide equal educational opportunities to all students, irrespective of place of residence or social background.

In the pre-reform system all students entered primary school (“kansakoulu”) at the age of seven. After four years in the primary school, at age 11, the students were faced with the choice of applying to general secondary school (“oppikoulu”) or continuing in the primary school. Admissions to the general secondary school track were based on an entrance examination, a teacher assessment and primary school grades. Those who were admitted to the general secondary school (52% of the cohort in 1970) continued first in the junior secondary schools for five years, and often went on to the upper secondary school for three additional years. At the end of the upper secondary school the students took the matriculation examination that provided eligibility to university-level studies.

Those who were not admitted or who did not apply to the general secondary school track continued in the primary school. The primary school lasted altogether eight years. Last two years of primary school concentrated on teaching vocational skills and were called continuation classes or “civic school”. These continuation classes that were compulsory for students that did not continue their schooling elsewhere had been introduced in the 1957 Primary School Act. After an amendment in 1963 municipalities could further extend these civic school courses by a year, thereby creating a nine-year primary school. The minimum school leaving age, regardless of the track, was sixteen, unless the student had already completed all required primary school courses. The pre-reform system is described schematically in the left-hand panel of Figure 1.

[FIGURE 1: SCHOOL SYSTEMS]

2.2 Content of the comprehensive school reform

The reform introduced a new curriculum and changed the structure of primary and secondary education. The new curriculum increased the academic content of education compared to the old primary school curriculum by increasing the share of mathematics and sciences. In addition, one foreign language became compulsory for all students. The new comprehensive school curriculum resembled the old general secondary school curriculum and exposed the

pupils who, in the absence of the reform would have stayed in the primary school, to a significantly more academic education.

The structure of the post-reform school system is described in the right-hand panel of Figure 1. Previous primary school, civic school and junior secondary school were replaced by a nine-year comprehensive school. At the same time the upper secondary school was separated from the junior secondary school into a distinct institution. After the reform, all pupils followed the same curriculum in the same establishments (comprehensive schools) up to age sixteen. The comprehensive school reform did not change the minimum school leaving age that had been sixteen ever since 1957.

After nine years in the new comprehensive school, all students could choose between applying to upper secondary school or to vocational schools. Admission to both tracks was based solely on comprehensive school grades.

2.3 The implementation of the comprehensive school reform

The implementation of the reform was preceded by a process of planning that lasted for two decades. Government working groups had proposed creating a comprehensive school already in 1948, 1957, 1959, and 1965. The first experimental comprehensive schools started their operation in 1967. Finally, in 1968 the Parliament approved the School Systems Act (467/1968) according to which the two track school system would be gradually replaced by a nine-year comprehensive school. The adoption of the new school system was to take place between 1972 and 1977, and the order in which the municipalities implemented the reform was to be loosely determined by geography starting from the Northern Finland where access to education was most limited. A regional implementation plan divided the country into six implementation regions and dictated when each region would implement the comprehensive school system. Regional school boards were created to oversee the transition process. The municipalities that were responsible of the school system could not select the reform date but were forced to follow the plan designed by the National Board of Education.

In each region, the five lowest primary school grades were to start in the comprehensive school immediately during the fall term of the year stated in the implementation plan. After this, each incoming cohort of first graders would start their schooling in the comprehensive

school. The pupils who were already above the fifth grade in the year that the region entered the reform would complete their schooling according to the pre-reform system. Thus, in each region it took approximately four years to complete the reform.

In addition to changes in the curriculum and in the timing of tracking, the reform also added a year of compulsory schooling for those students who, in the absence of the reform, would have chosen the less academic track in a municipality that only offered eight years of primary school. However, by the time the reform took place most municipalities already offered full nine years of primary school. Although the register data that used here do not include information on the actual years of schooling completed, the quinquennial censuses can be used to track the population share of 15-year olds who were enrolled in school. In 1975, when the reform had not yet reached the ninth grade, 92.6% of the fifteen-year-olds that would be on the ninth grade if they progressed at the normal speed were at school. This fraction remained at 92.6% in 1980 when the reform had reached the ninth grade in all but the last reform region. In 1985, well after the reform was completely implemented, the fraction of those turning fifteen still at school was only slightly higher, 93.9%. To us, these numbers suggest that the comprehensive school reform did not increase the minimum school leaving age. The effect of the reform can thus be interpreted as coming through changes in the curriculum and in the timing of tracking choices.

Figure 2 illustrates how the reform spread through the Finnish municipalities during 1972-1977. The municipalities assigned to the group where the reform was implemented in 1972 were predominantly situated in the northernmost province of Lapland. In 1973 the reform was implemented in the north-eastern regions. From thereon, the reform spread such that it was implemented in 1974 in the northwest, in 1975 in south-east, in 1976 in the south-west, and finally, in 1977 in the capital region of Helsinki.

[FIGURE 2: COMPREHENSIVE SCHOOL REFORM MAP]

2.4 The comprehensive school reform as a quasi-experiment

The Finnish comprehensive school reform is in many ways a promising natural experiment for evaluating the effects of tracking on student outcomes. A particularly useful setup was created by the regional implementation plan that dictated when each implementation region

moved into comprehensive school system. A fixed-effects approach allows for controlling for other simultaneous trends and regional differences and purging the estimate of school system from these confounding factors.

However, there are some caveats to the approach. First of all, as is clear from Figure 2, regional implementation plan did not strictly follow the geography. Some municipalities were placed into earlier or later implementation regions than their geographic neighbours. This suggests that, although the timing of the reform followed the national implementation plan, the allocation of municipalities to the reform year groups may not have been entirely random. Though the municipalities could not opt out from the plan or introduce comprehensive school before the assigned date, their views were heard before government ratified the implementation plan. It is not clear how much endogeneity in timing of the reform this hearing process implies. In any case we use a fixed effects approach that fully controls for the differences between the municipalities where the reform took place earlier and the municipalities where the reform took place later.

The comprehensive school reform also faced intensive resistance. Most common arguments against the reform were that abolishing tracking would reduce the quality of education. As a compromise, ability tracking was partially retained and some subjects therefore taught at different levels within the comprehensive school. In practice the reform students were divided into ability groups in foreign language and math classes, but studied all other subjects in their regular (not tracked) classes. This ability grouping was eventually abolished in 1985.

3. Data

A fundamental problem in assessing the effects of a school reforms on student performance is that students in separate school systems rarely participate in comparable tests. Sometimes it is possible to use nation-wide or international comparisons of student achievement. However, since most large-scale school reforms took place in 1960s and 1970s when testing was not as widespread as today, it is difficult to find tests implemented to representative and reasonably large samples of students from both pre- and post reform school systems.

This paper uses the Basic Skills test results from the Finnish Army. Since military service is mandatory in Finland, almost the entire male cohort takes the test. The Army test is given to

all new conscripts between the third and the sixth week of their service. The average age at the time when these men are tested is 20, so obviously also other factors than school system may have had an effect on the test results. On the other hand, the long-lasting outcomes of school systems are probably more interesting than the immediate effects on test results. In addition, the Basic Skills test is also a strong predictor of earnings and occupation later in life, so any effect of school system on the test scores will have important consequences for lifetime earnings.

The Finnish Army Basic Skills test is designed to measure general abilities. The Army uses the test results in selecting conscripts to officer training. The test consists of three subcategories: verbal, arithmetic, and logical reasoning. Each subtest includes forty multiple choice questions sorted in increasing order of difficulty. In the verbal reasoning subtest, the subject has to choose synonyms or antonyms of given words, select words that belong to the same category as the given word, exclude words from a group of words, and identify similar relationships between word pairs. The arithmetic reasoning test asks the subject to complete number series, solve verbally expressed mathematical problems, compute simple arithmetic operations, and choose similar relationships between pairs of numbers. The logical reasoning test is a standard “culture free” intelligence test based on Raven’s progressive matrices and its results should therefore be less affected by pre-test schooling.⁴ On the other hand, both the verbal and arithmetic reasoning categories test skills that are primarily taught in school.

The test was originally created in 1955 and re-designed in 1981. Exactly the same test was used over the entire time span analyzed here. From 1982 the test results are stored in the Army database that also includes personal identification numbers, making it possible to link the test results to information on test takers from other data registers. Our data include all conscripts who were born between 1962 and 1966 who were found in the Army database, i.e. those who started their military service after January 1982. There is some selectivity in the data due to the fact that it is possible to enter to military service as a volunteer before age 20. Thus some men in the oldest cohorts served before the Army register was created. Because of this selection problem with the older cohorts, the main analysis below will be conducted on the sample of conscripts born between 1964 and 1966. For these cohorts, it is also possible to observe the conscripts who volunteer for early service at ages 18 or 19.⁵ While it was not

⁴ The contents of the tests are described in detail in Tiihonen et al. (2005).

⁵ In the data, the average test scores are higher for those who serve as a volunteer compared to those who serve at age 20. To make the cohorts comparable, we could naturally control for the test age or limit the data to those

possible to extend the data to older cohorts as those data do not exist, it would have been possible to add more young cohorts to the data. However, since the cohorts born after 1966 would all be in the comprehensive school, adding younger cohorts would not be very informative.

It was possible to be exempted from the military service due to religious or ethical conviction though in 1980s this was rare. More common reasons for being exempt from military service were severe health conditions, most often related to mental health problems. However, even these criteria were substantially stricter in the 1980s than what they are today. A comparison of the number of observations by birth cohort in our data and the corresponding cohort size in the 1980 population census reveals that our test score data contain information on 85.3 percent of the relevant (1964-66) male birth cohorts⁶. This corresponds closely to the fraction of the cohort that served in the military in 1980s (Finnish Defense Command, 2000).

Figure 3 plots the distribution of the raw scores, i.e. the number of correct answers in each subtest. The distribution of the average score is plotted in the bottom right corner. Two features might be noted. First, there is plenty of variation in the test scores; the raw scores are distributed over the whole range from zero to forty. Second, the distribution of the test scores, particularly the distribution of the average score seems to be close to the normal distribution, a feature that will be exploited later.

FIGURE 3: DISTRIBUTION OF THE TEST SCORES]

Per our request, Statistics Finland linked the test scores from the Army data to Census data on Finnish population. The Statistics Finland longitudinal census file contains data on the entire population living in Finland in 1970, -75, -80, -85 and -90. From 1990 onwards information is available annually. Finnish Census data is almost entirely based on administrative registers. For example, information on education is based on the Register of Degrees and Examinations that collects data on all degrees granted directly by educational institutions. Data on the place

taking test at age 20 (75% of the sample). However, the reform may also have had an effect on the length of education and hence optimal timing of military service thereby making the test age an endogenous variable. Controlling for an endogenous variable or limiting the sample based on such variable would generally lead into biased results.

⁶ Our data is collected from the Finnish Army database and contains no information regarding those who did not serve in the military. Comparing the number of observations in our data to the overall cohort size in each region reveals slight variation in participation across cohorts and regions. It is also clear that those dismissed from the army due to health and mental reasons differ in many ways from the rest of the population. However, a simple regression where serving in the army is explained by the region and cohort dummies, and a comprehensive school dummy, indicates that the comprehensive school reform had no effect on the likelihood of serving in the military.

of residence in each census year is based on the Population Register. In general, these register data are of very high quality. Only a few persons have any missing data, and the main reasons for not being included in the census data are residing abroad and death. Therefore practically all conscripts were found in the register data and our data does not suffer from attrition problems that often plague similar studies.

From the census data information was gathered on the date of birth and the place of residence in 1970, -75 and -80, which jointly determine whether the individual attended a tracked or a comprehensive school system. Statistics Finland does not release these data with a municipality-identifier, but per our request created an indicator classifying municipalities into six categories according to the year in which the comprehensive school reform was implemented in each municipality. Except for those who moved between census years between municipalities that implemented the reform at different years, it is possible to accurately determine which school system was in place when the students were in the relevant age. The movers were dropped from the data used below, resulting in a reduction of the sample by 5%.⁷

The census data also include family codes that can be used to identify brother pairs and to gather information on parents' education and earnings. To be more exact, these family codes are based on persons living in the same household, not necessarily biological family members. We use the family codes from the 1975 census, when the oldest men in the sample were 13 years old and most likely still living at home.

Table 1 reports the mean test scores by cohort and reform region in the entire sample at disposal. The rectangular area in the bottom right-hand corner of the table shows the sample of cohorts born between 1964 and 1966 that will be used for the main analysis. In order to balance the size of the treatment and the control groups in this limited sample the regions that implemented the reform between 1972 and 1974 are excluded from the main analysis. This restriction of the sample also lessens the problem caused by the movers.⁸ If we had included regions where the reform was implemented in 1974 or before, 10% of the sample would have

⁷ As a robustness check we included the movers and determined the reform year based on the place of residence in the 1975 Census. This made practically no difference for the results.

⁸ Internal migration rates peaked in early 1970's and were about 25% higher in 1971-75 compared to 1976-80. In addition, municipality codes change because of the municipality mergers. Between 1971 and 1975 altogether 71 of 518 municipalities that had existed in 1970 merged with their neighbouring municipalities. Restructuring of local governments was much less intensive in late 1970s and hence the municipality codes that were used to determine the treatment status more stable between 1975 and 1980.

been lost because of the elimination of movers. This could already raise concerns of sample selectivity, particularly if one suspects that moving may be endogenous. As a robustness check, the results for the whole sample are reported in the Appendix.

As shown in Table 1, there are large differences across regions and a general increase in the test scores over time. These regional differences are correlated with the average education level and the average income in the region. An increase in the test scores over time, generally known as the Flynn-effect, has also been documented using same data by Koivunen (2007) for a longer time period. However, this naturally reflects also other differences between cohorts than those due to the school system.

The shaded area of the table indicates the students who attended the comprehensive school. Since these students are younger and concentrated in the regions with below average test scores, it is obvious that a cross-section comparison of regions or a time-series comparison of subsequent cohorts would not produce reliable estimates for the effect of the comprehensive school reform.

[TABLE 1: MEAN SCORE BY COHORT & REGION]

4. Estimation methods

Our goal is to estimate the causal effect of the school regime on the Army test scores. That is, to determine how an average student, or a student with certain characteristics, would have fared, had she or he been assigned to the reformed comprehensive system instead of the previous selective early tracking system. A fixed effects approach is used to control for regional differences as well as general trends over time. The effect of the comprehensive school reform is identified because the timing of the reform differs across regions.

Most of the estimates are based on the following regression model:

$$y_{ijt} = \alpha + \Omega'D_{ij} + \Psi'D_{it} + \beta C_{ijt} + \varepsilon_{ijt} \quad (1)$$

where y_{ijt} is the army test score of individual i who attended school in region j and belongs to cohort t . D_{ij} and D_{it} are region and cohort specific dummies, and C_{ijt} is an indicator for a pupil attending comprehensive school.

The parameter of interest in (1) is β . The identifying assumption is that the comprehensive school indicator, C_{ijt} , is uncorrelated with the error term conditional on the other regressors. This assumption, and the fact that D_{ij} and D_{it} enter (1) additively, reflects the basic differences-in-differences assumptions. Note, in particular, that we make no assumptions regarding the similarity of the regions where the reform took place early to those where reform took place later, nor do we claim that reform dates were randomly assigned. The parameter β is an unbiased estimate of the average causal effect of comprehensive schooling, if the timing of the reform is uncorrelated with other region-specific changes in student outcomes. In what follows, robustness of the results to these assumptions is also assessed by allowing for linear trends in test scores within regions

It is important to notice that regression (1) controls for the regional differences with six implementation region dummies but not for the full set of more than five hundred municipal fixed effects. The main reason is that we have no access to municipality codes due to data protection regulations. However, the only reason to include municipality dummies to regression (1) would be the concern that the reform took place earlier in non-randomly selected municipalities. But this is only a problem if the reform dummy is correlated with the municipality fixed effects. This correlation is fully absorbed by introducing the six implementation region fixed effects, since, within these regions, the implementation year does not vary across municipalities.

In addition to the mean effects of the reform, we also estimate (1) by interacting C_{ijt} with parental education and income. These results are informative in evaluating whether the reform was successful in improving the cognitive skills of students from disadvantaged family backgrounds. Furthermore, we also evaluate the effect of the reform on the variance of the test scores. A natural approach for examining this is the explicit modeling of the error variance of Equation (1). This could be performed in two steps by taking the residuals from Equation (1) and then explaining the squared residuals with the cohort and region dummies and the comprehensive school reform indicator. This procedure would be similar to standard heteroskedasticity tests. Moreover, since the OLS-residuals are unbiased estimates of the error terms, the two-step estimates are consistent.

A more straightforward approach is to model simultaneously the effect of the reform on both the mean and the variance of the test scores. Assuming that the error term follows a normal distribution, the test scores will be distributed as

$$y_{ijt} \sim \frac{1}{\sqrt{2\pi\sigma_{jt}^2}} \exp\left[-\frac{1}{2} \frac{(y_{ijt} - (\alpha + \Omega D_{ij} + \Psi D_{it} + \beta C_{ijt}))^2}{\sigma_{jt}^2}\right]. \quad (2)$$

The subscripts in σ_{jt}^2 indicate that the variance in the test scores may vary across regions and cohorts and may therefore be affected by the reform. The model is parameterized by assuming that log-variance is an additive function of the region, cohort and reform dummies. This also ensures that the variance estimates are positive and makes the coefficients easy to interpret.

$$\sigma_{jt}^2 = \exp(\alpha + \Gamma D_j + \Phi D_t + \gamma C_{jt}) \quad (3)$$

The log-likelihood function of the normal – heteroskedastic model is

$$\ln L = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^N (\alpha + \Gamma D_{ij} + \Phi D_{it} + \gamma C_{ijt}) - \frac{1}{2} \sum_{i=1}^N \left[\frac{(y_{ijt} - \alpha - \Omega D_{ij} - \Psi D_{it} - \beta C_{ijt})^2}{\exp(\alpha + \Gamma D_{ij} + \Phi D_{it} + \gamma C_{ijt})} \right] \quad (4)$$

where β measures the effect of the reform on the mean score and γ its effect on the variance.

5. Results

The baseline results are reported in Table 2. To facilitate the quantitative interpretation of the results, the test scores are converted into standard deviations units. Column (1) simply regresses the average test score on the comprehensive school dummy, and shows that those who attended the comprehensive school scored on average 0.07 standard deviations lower in the Army test. However, the results in Column (2) reveal that this negative correlation reflects the fact that regions with lower test scores implemented the reform first. When full sets of birth cohort and region dummies are included in the regression, the effect of comprehensive school is slightly positive and significant, at 0.03 standard deviations.

The causal interpretation of the result in column (2) of Table 2 relies on the standard differences-in-differences assumption that the changes in test scores in the reform regions would have been similar to the changes in the control regions in the absence of the reform. Given that the panel spans over several periods, this assumption can be relaxed somewhat by allowing for linear trends in test scores within regions. This is done in column (3), which allows for region specific linear trends. The results are not very sensitive to the inclusion of these trends. The effect of the reform is slightly larger but still relatively small at 0.05 standard deviation units.

Column (5) adds family fixed effects to the equation, thus identifying the effect of the reform from the differences between brothers that attended different school systems. The effect of the reform is not sensitive to the inclusion of family fixed effects but the estimates become very imprecise. Interestingly, adding family fixed effects also reverses the positive trend in the test scores, indicating that the birth order effect on the test scores is larger than the difference across the birth cohorts.⁹

[TABLE 2: BASIC RESULTS]

Table 3 examines separately the effect of the school reform on different tests. Column (1) regresses each test score separately on the region and cohort dummies and a dummy variable indicating whether the person had attended a comprehensive school. Column (2) again adds controls for the linear trends within regions and Column (3) controls for the family fixed effects. For brevity, only the coefficients of the comprehensive school dummy are reported in each case. According to the Table 3, the comprehensive school reform had no significant effects on either math or logical reasoning tests. The effect on the verbal ability test is positive. The size of the effect on verbal test scores is between 0.05 and 0.07 standard deviations units. Family fixed effect estimates tend to be much less precise than the estimates that exploit between-family variation, and are therefore never significantly different from zero or significantly different from the point estimates reported in Columns (1)-(2).

The finding that the comprehensive school reform has its largest effects on the verbal test was perhaps to be expected. After all, verbal skills are learned in schools, and hence the changes in school system may have effects on these skills. If indeed the logical reasoning test truly measures innate reasoning abilities, pre-test schooling should have little or no effect on the

⁹ The birth order effect was also found in a Norwegian study of the Army test scores (Kristenssen and Bjerkdal, 2007).

test. Finally, the changes in the mathematics teaching resulting from the reform were perhaps not as significant. As noted above, the ability grouping was retained in mathematics and, as a result, math classes continued to be taught at three different ability levels after the reform.

[Table 3: EFFECTS ON DIFFERENT TEST ITEMS]

Table 4 reports the maximum-likelihood estimates measuring the effects of the reform on both the mean and the (log)variance of the test scores. These equations are estimated separately for each test. All equations include cohort and region effects on both the mean and the variance, but only the effects of the comprehensive school are reported. The maximum-likelihood method produces very similar estimates on the effect of the reform on the mean scores as the linear regression model used in Tables 2 and 3. The effects are significant only for the verbal test and for the average score. The effects on the variance of the test scores are small. In the math test the effect is close to zero. In the verbal and logical reasoning test the reform reduced the variance between 2.5 and 3.2 percent. None of these effects, however, are statistically significant.

[Table 4: EFFECTS ON MEAN AND VARIANCE]

Tables 5A and 5B examine the effects of the comprehensive school reform by family background. Table 5A estimates regression models similar to those reported in Column (2) of Table 3 but adds an indicator of parents' education and its interaction with the reform dummy. Parents are classified as having higher education if at least one of the parents has completed at least 12 years of education. In the pre-reform schooling system this generally refers to a situation where the parent attended the more academic track. The parents' income is measured by summing the annual taxable income of both parents, deflating the income to the 1980 price level and taking an average over the census years 1970, -75 and -80. To facilitate the interpretation of the coefficients, parental earnings were normalized by subtracting the sample mean. This demeaning has no impact on the estimate of the interaction effect of the reform and parental earnings, but it makes it possible to interpret the main effect of the reform in table 5B as the effect of the reform on sons from families with sample mean income.

According to Table 5A, parental schooling has a clear effect on the test scores. Men with highly educated parents have a 0.31 standard deviations higher score in the verbal test, 0.32 standard deviations higher score in the math test, and 0.28 standard deviation higher score in the logical reasoning test. The effect of the reform – now referring to the effect on those with

less educated parents – is positive and statistically significant in all tests. The largest effect is again on the verbal test score, where the reform improved the test score of the conscripts with less educated parents by 0.08 standard deviations. This effect is sizeable, as it amounts to a quarter of the effect of parental education. More importantly, the interaction between parents' education and the comprehensive schooling is always negative and significant. A closer look at the average test score explains that the reform increased the score for those with less educated parents by about 0.07, but the interaction coefficient is approximately -0.05, producing a close to zero effect for men with highly educated parents.

Table 5B repeats the analysis using parents' income. The results are qualitatively similar to those using parents' education. Men with richer parents tend to score better in all tests, and the interaction between the parents' income and the reform dummy is negative for all tests.

[Tables 5A and 5B: EFFECTS BY FAMILY BACKGROUND]

As a final sensitivity check, we included a measure of “treatment intensity” that captures the number of years the students spent in the reformed school system. In the data that was used above, the students attended comprehensive school system for zero, five, six or seven years. In Table 6, these are all treated as separate dummy variables (omitting zero). While the estimated effects reach statistical significance only in the case of verbal test scores, the patterns of results is interesting. In general, the more years spent in the comprehensive school, the larger the effect of the reform.

[Table 6: EFFECTS BY YEARS IN THE COMPREHENSIVE SCHOOL]

6. Conclusions

Persistent differences in average test scores across countries and over time have received plenty of attention in recent years. One often suggested explanation for these differences is the educational system. In particular, the tracking of pupils into different groups by ability and aspirations has been considered a potentially important factor. However, both the economic theory and the available empirical evidence remain inconclusive when it comes to the effects of tracking regimes on test scores.

This paper estimated the effect of the comprehensive school reform on the Finnish Army Basic Skills Test scores. Unlike previous literature that had to rely on cross-country

comparisons or comparisons of regions within countries, here the effect of the comprehensive school reform on test scores is estimated using a differences-in-differences approach using single-country data. As such, the current study provides a more serious attempt at identifying the causal effect of school systems on test outcomes.

On average, the reform had a small positive effect on the average verbal test scores and no significant positive or negative effect on the average arithmetic or logical reasoning test results. Most interestingly, in all the tests, the effect of the reform was positive and significant in families where the parents had only basic education or low income. The effect on the verbal test score was significant in magnitude, as it was comparable to about a fourth of the effect of parental education. In verbal test score, the effect of the reform on sons with less educated parents was also considerable. Finally, the effect of the reform on the variance of the test scores was not significant.

Our results are only partially in line with earlier studies by Hanushek and Wößmann (2006) and Brunello and Checchi (2007). We find no significant effects on the variance of the test scores but find that comprehensive school has beneficial effects on mean performance in some tests. More importantly, we show that the postponement of tracking and exposing children to a more academic curriculum for a longer period appears reduce test score differences across socio-economic groups. This also indicates that the positive effects of school reforms on social mobility demonstrated by e.g. Meghir and Palme (2005) as well as Pekkarinen et al. (2009) are partially explained by differential effects of the reform on skill formation in different socio-economic groups.

References

- Brunello, G. and D. Checchi (2007), "Does school tracking affect equality of opportunity? New international evidence", *Economic Policy*, Oct 2007, pp. 781-861.
- Duflo, E., P. Dupas and M. Kremer (2008), "*Peer effects and the impact of tracking: Evidence from a randomized evaluation in Kenya*", NBER Working Paper, 14475.
- Finnish Defense Command (2000) "*The Finnish Defense Forces, Annual Report 2000*" (in Finnish), Defense Command Public Information Division, Helsinki.
- Galindo-Rueda, F. and A. Vignoles (2005), "*The heterogeneous effect of selection in secondary schools: Understanding the changing role of ability*," CEE Discussion Paper 52.
- Hanushek, E. and L. Wößmann (2006), "Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries," *Economic Journal* vol. 116, pp. C63-C76.
- Hanushek, E. and L. Wößmann (2009), "The role of cognitive skills in economic development", *Journal of Economic Literature*, 46: 3, 607-668.
- Kerckhoff, A., K. Fogelman, D. Crook and D. Reeder (1996), "*Going Comprehensive in England and Wales. A Study of Uneven Change*", London: The Woburn Press.
- Koivunen, S. (2007), "Suomalaismiesten kognitiivisen kykyprofiilin muutokset 1988-2001. Flynnin efektiä suomalaisessa aineistossa?" [Changes in cognitive skill profile among Finnish men. Flynn effect in Finnish data?], Master's thesis (in Finnish), University of Jyväskylä.
- Kristensen P. and T. Bjerkedal (2007), "Explaining the Relation Between Birth Order and Intelligence", *Science*, vol. 316 (5832), p. 1717.
- Lazear, E. (2001), "Educational Production", *Quarterly Journal of Economics* vol. 116 (3), pp. 777-803.
- Manning, A. and J. Pischke (2006), "*Comprehensive versus selective schooling in England in Wales: What do we know?*" NBER Working Paper No. 12176.
- Meghir, C. and M. Palme (2005), "Educational reform, ability, and parental background", *American Economic Review* vol. 95 (1), pp. 414-424.

OECD, (2003): *“Learning for Tomorrow’s World: First Results from PISA 2003”*, OECD Publications, Paris.

Pekkarinen, T., R. Uusitalo and S. Pekkala (2009), “Education policy and intergenerational income mobility: Evidence from the Finnish comprehensive school reform”, *Journal of Public Economics* 93, 965 - 973.

Statistics Finland (1986), *“Structure of population and vital statistics”*, Official Statistics of Finland VI A: 150, Central Statistical Office of Finland, Helsinki 1986.

Tiihonen, J., J. Haukka, M. Henriksson, M. Cannon, T. Kiesepä, I. Laaksonen, J. Sinivuo, and J. Lönnqvist, (2005), “Premorbid intellectual functioning in bipolar disorder and schizophrenia: Results from a cohort study of male conscripts”, *American Journal of Psychiatry* vol. 162, pp. 1904-1910.

Waldinger, F. (2006), “Does tracking affect the importance of family background on students' test scores?” mimeo, London School of Economics.

Figure 1 Finnish school systems before and after the comprehensive school reform

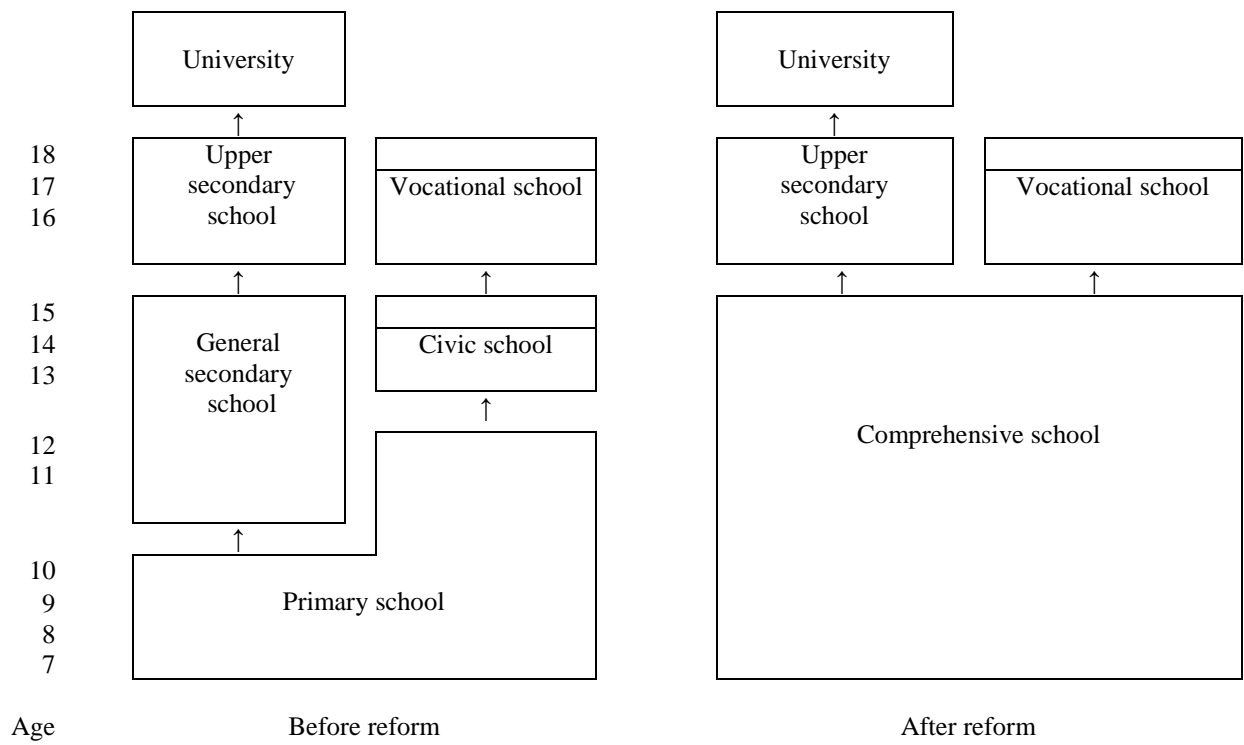


Figure 2 The implementation of the comprehensive school reform across regions 1972-1977

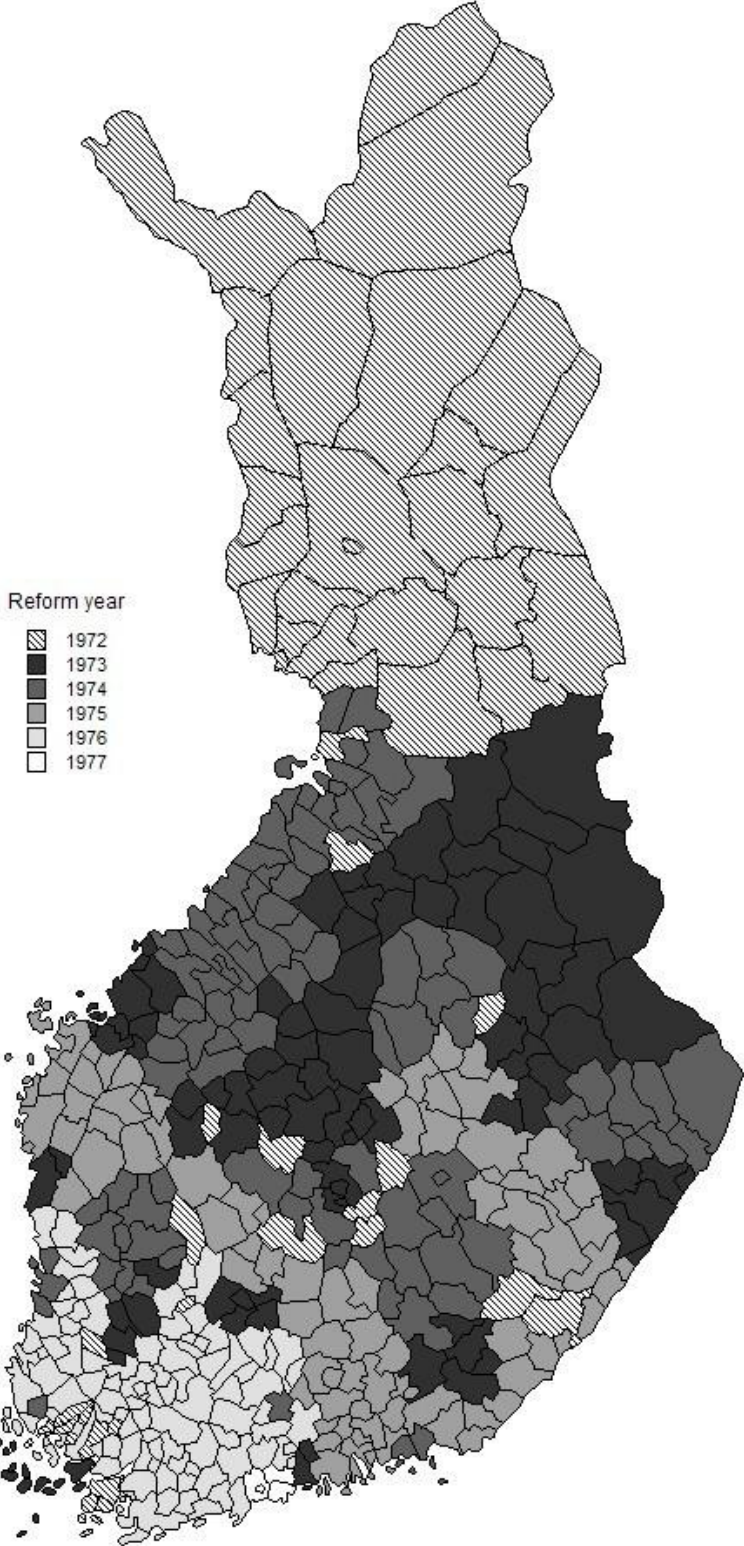


Figure 3 Distribution of the test scores

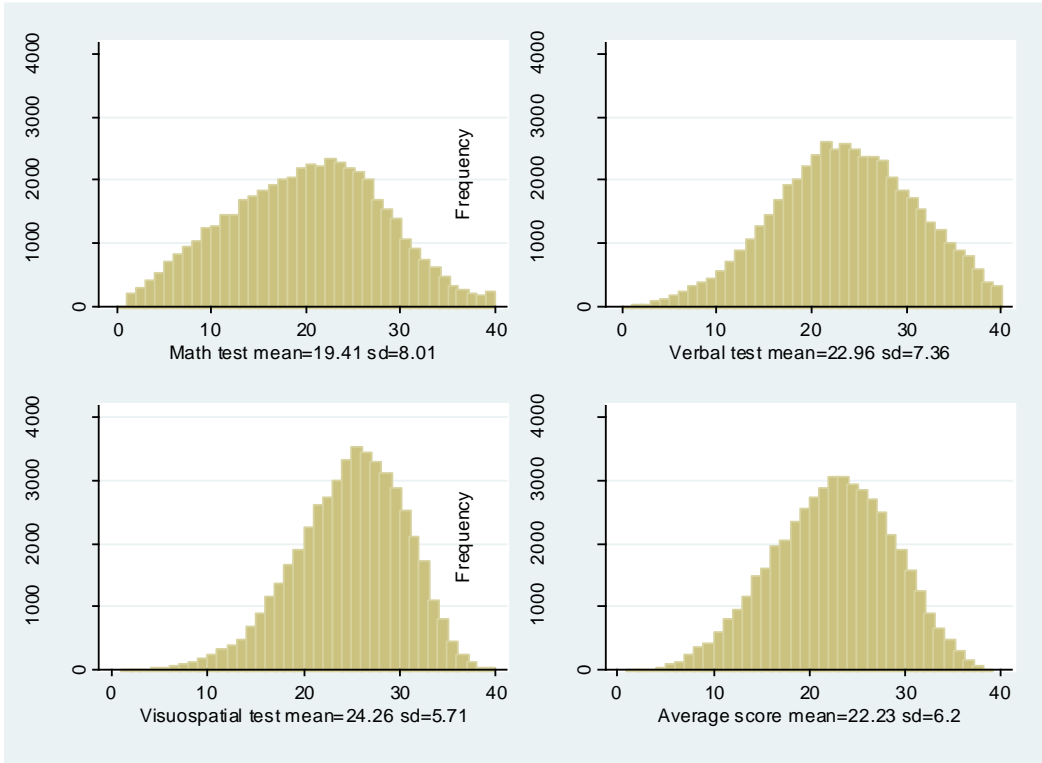


Table 1 Average test score by region and cohort

Birth cohort	Reform year						Total
	1972	1973	1974	1975	1976	1977	
1962	19.96 [2,471]	19.87 [3,629]	20.30 [5,286]	20.73 [5,095]	21.26 [5,304]	22.57 [2,822]	20.76 [24,607]
1963	20.82 [2,776]	20.57 [4,126]	20.94 [6,037]	21.35 [6,190]	21.91 [6,174]	22.96 [3,516]	21.42 [28,819]
1964	20.70 [2,663]	20.4 [3,967]	21.04 [5,751]	21.44 [5,964]	21.85 [6,188]	23.07 [3,696]	21.44 [28,229]
1965	20.92 [2,494]	20.74 [3,706]	21.09 [5,517]	21.48 [5,846]	22.18 [5,790]	23.02 [3,594]	21.60 [26,947]
1966	21.42 [2,023]	21.13 [3,037]	21.55 [4,732]	21.74 [5,176]	22.35 [5,094]	23.41 [3,344]	21.96 [23,406]
Total	20.74 [12,427]	20.52 [18,465]	20.97 [27,323]	21.35 [28,271]	21.91 [28,550]	23.02 [16,972]	21.44 [132,008]

Note: The shaded areas indicate cohorts that were affected by the post-reform educational system. Number of observations in each cell is in square brackets below the mean score in each cell.

Table 2 Effects on test score average: Cohorts 1964-1966 and regions 1975-1977

	(1) No controls	(2) Region & cohort	(3) Regional trends	(5) Family fixed effects
Reform	-0.0727 (0.0771)	0.0312* (0.0162)	0.0451* (0.0238)	0.0328 (0.0461)
Birth year 1965		0.00637 (0.0123)	0.00919 (0.0127)	-0.115*** (0.0361)
Birth year 1966		0.0413*** (0.0148)	0.0493*** (0.0176)	-0.110*** (0.0394)
Reform region 1976		0.0998*** (0.0115)	0.114*** (0.0253)	
Reform region 1977		0.261*** (0.0160)	0.290*** (0.0330)	
Linear trend region 1976			-0.00947 (0.0174)	
Linear trend region 1977			-0.0197 (0.0184)	
Constant	3.653*** 51920	3.465*** 51920	3.448*** 51920	3.652*** 51920
R-squared	0.001	0.009	0.009	0.006

Notes: The dependent variable is an unweighted average in three tests scaled into standard deviation units. Standard errors are clustered at the implementation region level.

Table 3 Effects in different tests

	(2) Region & cohort	(3) Regional trends	(5) Family fixed effects
Math test	0.0135 (0.0162)	0.0237 (0.0239)	0.0121 (0.0479)
Verbal test	0.0510*** (0.0162)	0.0742*** (0.0239)	0.0684 (0.0472)
Logical reasoning	0.0238 (0.0162)	0.0246 (0.0238)	0.0222 (0.0521)

The entries in the table are coefficients of the dummy variable indicating that the person attended comprehensive school. Each regression model is estimated separately and includes cohort and region fixed effects. Standard errors are clustered at the implementation region level.

Table 4 The effect of the reform on mean and variance of the test scores

	(1) Math	(2) Verbal	(3) Logical reasoning	(4) Average score
OLS estimates				
Effect on mean	0.0135 (0.0162)	0.0510*** (0.0162)	0.0238 (0.0162)	0.0312* (0.0162)
ML estimates				
Effect on mean	0.0136 (0.0162)	0.0513*** (0.0162)	0.0241 (0.0161)	0.0315* (0.0162)
Effect on log variance	0.00974 (0.0230)	-0.0249 (0.0230)	-0.0319 (0.0229)	-0.0182 (0.0230)

The entries in the table are coefficients of the dummy variable indicating that the person attended comprehensive school. Each regression model is estimated separately and includes cohort and region fixed effects.

Table 5A: Effect of the reform by parents' education

	(1) Math	(2) Verbal	(3) Logical Reasoning	(4) Average score
High ed. parents	0.318*** (0.0176)	0.309*** (0.0176)	0.284*** (0.0176)	0.348*** (0.0176)
Reform	0.0429** (0.0212)	0.0848*** (0.0212)	0.0621*** (0.0212)	0.0705*** (0.0211)
Reform × high ed. parents	-0.0383* (0.0208)	-0.0392* (0.0208)	-0.0458** (0.0208)	-0.0485** (0.0207)
Constant	2.178***	2.851***	3.980***	3.291***
Observations	46065	46135	46139	46057
R-squared	0.028	0.025	0.025	0.033

Table 5B: Effect of the reform by parents' income

	(1)	(2)	(3)	(4)
Log parents' income	0.390*** (0.0175)	0.351*** (0.0175)	0.310*** (0.0175)	0.400*** (0.0175)
Reform	0.0205 (0.0170)	0.0608*** (0.0171)	0.0329* (0.0171)	0.0411** (0.0170)
Reform × Log parents' income	-0.0446** (0.0202)	-0.0283 (0.0203)	-0.0121 (0.0203)	-0.0332* (0.0202)
Constant	2.394***	3.060***	4.174***	3.527***
Observations	46041	46111	46115	46033
R-squared	0.038	0.033	0.032	0.043

Parents' income is average log income of parents from the 1970, -75 and -80 census data inflated to the 1980 price-level using the consumer price index. Deviation from the mean log parents' income was used so that the effect of the reform can be interpreted as the effect at the mean income level.

Table 6 The effect of the reform by treatment intensity

	(1)	(2)	(3)	(4)
	Math	Verbal	Logical	Average score
5 years in comprehensive school	0.0250 (0.0274)	0.0874*** (0.0273)	0.0382 (0.0273)	0.0569** (0.0273)
6 years in comprehensive school	0.0353 (0.0459)	0.118** (0.0459)	0.0715 (0.0458)	0.0859* (0.0459)
7 years in comprehensive school	0.0485 (0.0691)	0.162** (0.0691)	0.0665 (0.0690)	0.109 (0.0691)
Constant	2.321***	2.963***	4.107***	3.435***
Observations	51932	52011	52017	51920

Reform indicator is decomposed into three dummies depending on the amount of years that individuals spent in the comprehensive school. The regression uses the same set of controls as in column 2 of table 3. Standard errors are clustered at the implementation region level.

Appendix:

Table A1: Reform effects using full sample

VARIABLES	(1) Math	(2) Verbal	(3) Logical Reasoning	(4) Average score
Reform	0.0142 (0.0110)	0.0424*** (0.0111)	0.0104 (0.0112)	0.0243** (0.0110)
Constant	1.584***	2.271***	3.445***	2.648***
Observations	142051	142286	142322	142001
R-squared	0.070	0.058	0.047	0.074

In all columns the model includes a full set of dummy variables for region and cohort and a set of dummy variables for age on the test date. Regressions also control for linear regional trends in test scores. Standard errors are clustered at the implementation region level.

Table A2a: Effect of the reform by parents' education using full sample

	(1) Math	(2) Verbal	(3) Logical Reasoning	(4) Average score
High ed. parents	0.252*** (0.00919)	0.244*** (0.00926)	0.231*** (0.00931)	0.276*** (0.00916)
Reform	0.0225** (0.0112)	0.0410*** (0.0113)	0.0214* (0.0113)	0.0311*** (0.0112)
Reform × high ed. parents	-0.0352*** (0.0113)	-0.0285** (0.0114)	-0.0268** (0.0115)	-0.0356*** (0.0113)
Constant	1.483***	2.217***	3.358***	2.561***
Observations	127206	127404	127432	127167
R-squared	0.085	0.072	0.059	0.092

Table A2b Effect of the reform by parents' income using full sample

	(1) Math	(2) Verbal	(3) Logical Reasoning	(4) Average score
Log parental income	0.309*** (0.00874)	0.281*** (0.00881)	0.276*** (0.00886)	0.326*** (0.00871)
Reform	0.00498 (0.00944)	0.0271*** (0.00952)	0.00803 (0.00957)	0.0136 (0.00940)
Reform × Log parental income	-0.0345*** (0.0106)	-0.0283*** (0.0107)	-0.0288*** (0.0107)	-0.0337*** (0.0105)
Constant	1.653***	2.380***	3.514***	2.744***
Observations	127120	127318	127346	127081
R-squared	0.094	0.079	0.066	0.102