Discussion Papers

School tracking and development of cognitive skills – additional results

Sari Pekkala Kerr
Wellesley College

Tuomas Pekkarinen
Aalto University, IZA and IFAU

Roope Uusitalo
HECER, IZA and IFAU

Discussion Paper No. 350
August 2012

ISSN 1795-0562

HECER – Helsinki Center of Economic Research, P.O. Box 17 (Arkadiankatu 7), FI-00014
University of Helsinki, FINLAND, Tel +358-9-191-28780, Fax +358-9-191-28781,
E-mail info-hecer@helsinki.fi, Internet www.hecer.fi

HECER
Discussion Paper No. 350

School tracking and development of cognitive skills – additional results*

Abstract

This paper evaluates the effects of selective vs. comprehensive school systems on military test scores in mathematical, verbal and logical reasoning skills tests. We use data from the Finnish comprehensive school reform which replaced the old two-track school system with a uniform nine-year comprehensive school. The paper uses a differences-in-differences approach and exploits the fact that the reform was implemented gradually across the country during a six-year period. 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 scores on all tests for the students whose parents had only basic education.

JEL Classification: H52, I21

Keywords: education, school system, tracking, comprehensive school, test scores.

Sari Pekkala Kerr
Wellesley College
WCW, 106 Central Street
Wellesley, MA 02481
USA
e-mail: skerr3@wellesley.edu

Tuomas Pekkarinen
Aalto University School of Economics
P.O. Box 21240
00076 Aalto
FINLAND
e-mail: tuomas.pekkarinen@aalto.fi

Roope Uusitalo
Helsinki Center of Economic Research (HECER)
P.O. Box 17 (Arkadiankatu 7)
FI-00014 University of Helsinki
FINLAND
e-mail: roope.uusitalo@helsinki.fi

* This working paper contains additional results that did not fit into the version submitted to Journal of Labor Economics. The authors would like to thank David Autor, Alexis Brownell, Lidia Farré, William Kerr, Sandra McNally, Eva Mörk, as well as seminar participants at London School of Economics, Universitat de Alicante, Wellesley College, ESPE Conference in London, and EALE Conference in Amsterdam for helpful comments. Pekkarinen and Pekkala Kerr are grateful for financial assistance from the Academy of Finland and Yrjö Jahnsson Foundation, respectively.

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, the variance of achievement scores is the extent and timing of tracking, or ability grouping of students. For example, the OECD has repeatedly argued that variation in student performance tends to be higher in countries with early tracking policies (e.g. OECD, 2003). High variance of student achievement and its correlation with family background have been seen as problematic from the perspective of equality of opportunity. On the other hand, postponing tracking could lower the quality of teaching at least for the high-ability students. Implicit in this debate is an efficiency-equity trade-off: postponement of tracking could improve equality but might decrease the average achievement.

The timing of tracking differs significantly between comprehensive and selective school systems. 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.

In this study we evaluate the effect of the Finnish comprehensive school reform on cognitive skills tests. Finland had a selective two-track school system until the 1970's, when the school reform replaced the old two-track system with a uniform comprehensive school system that is similar to those 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, 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 previously unavailable data from the Finnish Defense Forces we can partially open the 'black box' relating school systems to labor market outcomes. We use scores from cognitive skills tests taken at the beginning of mandatory military service. These data allow us to examine how the comprehensive school system affected arithmetic and verbal skills as well as logical reasoning ability. This is particularly important as recent research emphasizes the role of cognitive skills, rather than mere school attainment, as important determinant of individual earnings, the distribution of income and economic growth. (Hanushek and Wößmann, 2008). Yet existing evidence on the impact of major school reforms on cognitive skills is still scarce.

We use a similar differences-in-differences approach as in Pekkarinen et al. (2009) and exploit the fact that the school reform was implemented gradually across regions. However, we evaluate the effects of the school reform directly on the distribution of skills that the students are supposed to learn in school. Our results show that the reform had a small positive effect on verbal test scores, but little effect on the mean performance in the arithmetic or logical reasoning tests. However, the reform significantly improved the scores in all tests for students whose parents had only basic education or low income. At the same time, the reform had no negative effects on the test scores of students from more advantaged backgrounds. These results are qualitatively in line with the results in Pekkarinen et al. (2009) where it was shown that the comprehensive school reform had a substantial negative effect on the intergenerational income elasticity. However, we find that the effects on cognitive skills are far too small to fully explain the effects on income.

The rest of the paper proceeds as follows. In the next section, we review the literature on the effects of school tracking and compare our approach to the previous studies. We then describe the content and the implementation of the Finnish comprehensive reform. The fourth section describes the data and the Finnish Army Basic Skills Test, the results of which 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 in section five and in the sixth section we discuss the results. The seventh section concludes.

2. Previous literature

Economic theory provides somewhat ambiguous predictions on the effect of comprehensive versus selective school systems on student achievement. A comprehensive system, where the entire cohort is in the same class, increases heterogeneity in the classroom. This probably makes classes more difficult to teach and thereby may lower student achievement (Lazear, 2001). However, any changes in the class composition may also affect student achievement due to peer effects. 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.[1]

Even if the effect on the average student achievement is ambiguous, a comprehensive school system should decrease the variance of test scores. This is due both to a more homogenous curriculum and to less segregated peer groups. Furthermore, as early educational choices are more likely to be determined by family background (Brunello and Checchi, 2007), later tracking age in the comprehensive system may reduce the correlation between the test scores and the family background.

The most convincing empirical evidence on the effects of ability tracking on test scores comes from a field experiment in Kenya where randomly selected schools implemented tracking and non-tracking policies. Duflo et al. (2011) 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 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. In addition, the debate on the relative advantages of different school systems is primarily concerned with 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) find that the variance in test scores in international student

[1]In the Lazear (2001) model the peer effects cannot be distinguished from the curriculum effects. In his model ill-behaving students stop the teaching for the entire class. Hence, average student quality determines how much time the teacher can spend in teaching and how much material can be covered.

assessments 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. In contrast, Brunello and Checchi (2007) and Waldinger (2006), who use a similar cross-country approach, find no effects on the test score variance. These conflicting results from previous studies reflect, in part, the difficulties in analyzing the effects of school system based on cross-country data. While these studies try to control for variation due to other factors by controlling for early test scores (Hanushek and Wößmann, 2006; 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 are reliably accounted for.

Analyzing 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) and 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. 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 disentangle the effect of school systems from the regional differences using common data sources such as the National Child Development Survey, that only contain a single cohort.

Relative to the earlier studies, the distinct advantage of the Finnish reform is the availability of panel data from several cohorts, which avoids the need to rely exclusively on cross-sectional variation. The Finnish comprehensive school reform was implemented gradually region by region between 1972 and 1977. This gradual implementation allows us to control 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 differed in age.

3. Comprehensive school reform[2]

3.1 Background

Finland introduced a wide-ranging comprehensive school reform in the 1970's. Similar reforms had already taken place in Sweden in the 1950s and in Norway in the 1960s (Meghir and Palme, 2005; Aakvik et al. 2010). The Finnish comprehensive school reform abolished the old two-track school system and created a uniform nine-year comprehensive school. The main 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 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 for 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. The last two years of primary school concentrated on teaching vocational skills and were called continuation classes or "civic school". 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] This section draws on Pekkarinen et al. (2009).

3.2 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. The previous primary schools, civic schools and junior secondary schools were replaced by nine-year comprehensive schools. 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. After nine years in the new comprehensive school, students could choose between applying to upper secondary school or to vocational schools. Admission to both tracks was based solely on comprehensive school grades.

Unlike comprehensive school reforms in many other European countries, the Finnish reform did not extend the length of compulsory schooling or change the minimum school leaving age, which had been sixteen ever since 1957. Although the pre-reform system obliged the municipalities to provide only eight years of primary school, analysis of quinquennial census data reveals that by the time of the comprehensive school reform most municipalities were already offering a full nine years of primary school for pupils who did not go to junior secondary school. In 1975, when the reform had not yet reached the ninth grade, 92.6% of the fifteen-year-olds that would be in the ninth grade if they progressed at the typical speed were still in 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 while still in school was only slightly higher, 93.9%. To us, these numbers suggest that the comprehensive school reform did not increase the actual 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, which naturally also implies changes in peer groups.

3.3 The implementation of the comprehensive school reform

The implementation of the reform was preceded by a process of planning that lasted for two decades. The first experimental comprehensive schools started their operation in 1967. 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 system. The adoption of the new school system was to take place between 1972 and 1977. 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 for operating 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, pupils in the five lowest primary school grades started in the comprehensive school during the fall term of the year stated in the implementation plan. After that, each incoming cohort of first graders started their schooling in the comprehensive school. The pupils who were already above the fifth grade in the year when the reform was implemented in their home region completed their schooling according to the pre-reform system. Thus, in each region it took approximately four years to fully complete the reform.

Figure 2 illustrates how the reform spread through the Finnish municipalities during 1972 to 1977. The municipalities in which 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, in 1974 in the northwest, in 1975 in the south-east, in 1976 in the south-west, and finally, in 1977 in the capital region of Helsinki.

[Figure 2: COMPREHENSIVE SCHOOL REFORM MAP]

The comprehensive school reform faced intense resistance. The most common argument against the reform was that abolishing tracking would reduce the quality of education. As a compromise, ability tracking was partially retained so that math and foreign languages were taught at different levels within the comprehensive school. This ability grouping was eventually abolished in 1985.

A reform of this scale naturally implied important changes in teacher education and the internal organization of schools. In the old system primary school teachers were educated in

separate non-university institutions. Initially, these teachers continued to teach in the new comprehensive schools. Eventually, teacher education was moved into newly founded schools of education within universities. Over time the reform therefore led to an increase in the average duration and an improvement in the quality of teacher training.

The implementation of the Finnish comprehensive school reform makes it 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 municipality moved into the comprehensive school system. This allows us to use a fixed-effects approach to control for the changes over time and any regional differences between the municipalities that were assigned to different implementation regions.

Given that we use data from the very first cohorts that were affected by the reform, the effects are likely to be somewhat attenuated. For example, in the early years ability tracking was retained in certain subjects, most teachers still only had the shorter primary school teacher education, and the merging of separate schools into the same physical units probably caused disruptions in the organizational structure.

4. Data

A fundamental problem in assessing the effects of school reform on student performance is that students in separate school systems rarely participate in comparable tests. Sometimes it is possible to use nation-wide performance evaluations 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 administered to representative and reasonably large samples of students from both pre- and post-reform school systems.

This paper uses the results from the Basic Skills test of the Finnish Army. Since military service is mandatory for men in Finland, almost the entire male cohort takes the test. The average age at time of testing is 20, so obviously factors other than the school system may also have had an effect on the test results. On the other hand, the longer-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.[3]

The Finnish Army Basic Skills test is designed to measure general abilities. The Army uses these test results in selecting conscripts to officer training. Unlike the Swedish and Norwegian military test score data, used for example by Black, Devereux, and Salvanes (2010) and by Lindqvist and Vestman (2011), the Finnish test score data are available in a disaggregated form. The test consists of three subcategories: verbal, arithmetic, and logical reasoning. Each subtest includes forty multiple choice questions. 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 a 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 same test was used over the entire time span analyzed here. From 1982 onward, 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 and 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 military service as a volunteer before age 20. Thus some men in the oldest cohorts served before the Army register was created. We experimented with several solutions to this problem. For example, we limited the analysis to those who served in the army at age 20. We also restricted the data to men born between 1964 and 1966, i.e., those that we can observe even if they volunteered for early service at age 18 or 19. Since this made qualitatively little difference, we only report the results from using the full sample and simply control for age at test.[5]

[3] Uusitalo (1999) reports that recruits who scored one standard deviation higher in the Basic Skills tests earn on average 6% more than recruits with similar education and experience but lower test scores.
[4] The contents of the tests are described in detail in Tiihonen et al. (2005).
[5] Results based on restricted samples can be found in Appendix 3.

It is 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 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 male birth cohorts[6]. This corresponds closely to the reported fraction of the cohort that served in the military in the 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. These histograms clearly show that there is plenty of variation in the test scores; the raw scores are distributed over the whole range from zero to forty. Also, the distribution of the average scores, is symmetric and not very far from normal distribution.

[Figure 3: DISTRIBUTION OF THE TEST SCORES]

Per our request, Statistics Finland linked the test scores to census data on the 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 are based almost entirely on administrative registers. For example, information on the place 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.

Census data were used to gather information on the pupils' date of birth and place of residence in 1970, -75 and -80, which jointly determine whether the individual attended a tracked or a comprehensive school. 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

[6]Our data are collected from the Finnish Army database and contain no information on those who did not serve in the military. 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 comparison of the number of observations per cohort and region in our data to the overall cohort size in each region indicates that the comprehensive school reform had no effect on the likelihood of serving in the military.

implemented in each. Except for those who moved between census years between two municipalities that implemented the reform at different years, it is possible to accurately determine which school system was in place when the students were at the relevant age. The movers were dropped from the data used below, resulting in a reduction of the sample by 10%.[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 thirteen years old and most likely still living at home.

Table 1 reports the mean test scores by cohort and reform implementation region as deviations from overall sample means in standard deviation units. Implementation regions are defined by the year when the reform took place in the region. There are large differences across regions and a general increase in the test scores over time. These regional differences are correlated with the average parental education and income levels of the regions, reported in the last two rows of the table. An increase in test scores over time, generally known as the Flynn effect, has also been documented using the same data by Koivunen (2007) for a longer time period. However, this effect also reflects differences between cohorts other than those due to the school system.

The shaded area of the table indicates the students who attended comprehensive school. Since these students are younger and are 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. Similar tabulations are shown for the three subtests in Appendix 1.

[Table 1: MEAN SCORE BY COHORT & REGION]

[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.

5. 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:

$$y_i = \alpha + \Omega' D_{j(i)} + \Psi' D_{t(i)} + \delta A_i + \beta C_{j(i)t(i)} + \varepsilon_i \qquad (1)$$

where y_i is the army test score of individual i, attending school in region j and belonging to cohort t. D_j and D_t are region and cohort specific dummies and A_i is the age at test. $C_{j(i)t(i)}$ is an indicator, varying across cohorts and regions, that pupil i attended comprehensive school.

The parameter of interest in (1) is β. The identifying assumption is that the comprehensive school indicator, $C_{j(i)t(i)}$, is uncorrelated with the error term conditional on the other regressors. This assumption, and the fact that D_j and D_t enter (1) additively, reflect 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.

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 the municipality codes due to data protection regulations. However, the only reason to include municipality dummies in 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.

The fact that we only have 5 (cohorts) x 6 (regions) = 30 observations for identification purposes obviously has implications for statistical inference. We deal with this by clustering the standard errors at region level using the Moulton-procedure programmed for Stata by Angrist and Pischke (2009). Following Cameron, Miller, and Gelbach (2007), we also make a small sample correction for the standard errors by magnifying the residuals by $\sqrt{G/(G-1)}$, where *G* is the number of regions, and use critical values from a t-distribution with *G-2* degrees of freedom.[8] To facilitate the interpretation of the statistical significance of the results with non-standard critical values, we report the 95% confidence intervals instead of standard errors of our regression coefficients in the tables.

We also estimate (1) by interacting $C_{j(i)t(i)}$ with parental education and income. These results are informative in evaluating whether the reform was particularly successful in improving the cognitive skills of students from disadvantaged family backgrounds relative to other students.

Furthermore, we also evaluate the effect of the reform on the variance of the test scores. A straightforward approach for examining this 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

$$y_i \sim \frac{1}{\sqrt{2\pi\sigma^2_{j(i)t(i)}}} \exp\left[-\frac{1}{2} \frac{\left(y_i - (\alpha + \Omega D_{j(i)} + \Psi D_{t(i)} + \delta A_i + \beta C_{j(i)t(i)})\right)^2}{\sigma^2_{j(i)t(i)}} \right] \quad (2)$$

The subscripts in σ^2_{jt} 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.

[8]We were also concerned that serial correlation within regions could lead to a bias in the standard errors. However, calculating mean residuals by region and cohort from equation 1 and regressing these mean residuals on the lagged residuals indicates that the first order autocorrelations of the residuals in different test items are between -.09 and .11. These estimates are consistent with the null of no autocorrelation according to an autocorrelation test suggested by Wooldridge (2002), i.e. regressing the first differenced residuals on their lagged values and testing whether the resulting coefficient is equal to -.5. As a final check we also estimated the confidence intervals using a wild bootstrap-t procedure that retains the cluster structure in the data and, according to Monte Carlo results by Cameron et al. (2008), performs well even with only 6 clusters. In general, the width of the confidence intervals does not vary much across the different cluster correction procedures, nor are the cluster corrected standard errors very different from the OLS standard errors once cohort and region fixed effects are included in the model. We report alternative estimates of confidence intervals of key parameters in Appendix 2, but our main conclusion is that, given the lack of significant autocorrelation in the data, the choice of cluster correction method is not very important. This finding is consistent with the Monte Carlo experiments in Bertrand et al. (2004).

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

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

$$\ln L = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\left(\alpha + \Gamma D_{j(i)} + \Phi D_{t(i)} + \phi A_i + \gamma C_{j(i)t(i)}\right) - \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\left(y_i - \alpha - \Omega D_{j(i)} - \Psi D_{t(i)} - \delta A_i - \beta C_{j(i)t(i)}\right)^2}{\exp\left(\alpha + \Gamma D_{j(i)} + \Phi D_{t(i)} + \phi A_i + \gamma C_{j(i)t(i)}\right)}\right] \qquad (4)$$

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

6. Results

6.1 Average effects

The baseline results are reported in Table 2. To facilitate the quantitative interpretation of the results, the test scores are converted into standard deviation units. Column (1) simply regresses the average test score on the comprehensive school dummy, and shows that those who attended comprehensive school scored on average 0.095 standard deviations lower in the Army test. However, the results in column (2) reveal that this negative correlation reflects the fact that regions with on-average lower test scores implemented the reform first. When full sets of birth cohort and region dummies are included in the regression, this negative correlation is removed and we fail to find any significant effect of comprehensive school on average test scores.

The causal interpretation of the result in column (2) of table 2 relies on the standard difference-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 several periods, this assumption can be relaxed somewhat by adding region-specific linear trends in test scores as we do in column (3). The effect of the reform is now positive and close to being statistically significant at 5% level but very small at 0.025 standard deviation units.

Column (4) 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 estimates become imprecise but are still very close to those reported in column (3). 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 the effect of the school reform on different tests in turn. Column (1) regresses each test score separately on the region and cohort dummies and a dummy variable indicating whether the person attended a comprehensive school. Column (2) again adds separate region-specific linear trends and column (3) controls for the family fixed effects. For brevity, only the coefficients of the comprehensive school dummy are reported in each column. The comprehensive school reform had no significant effects on either math or logical reasoning tests. The effect on the verbal ability test is positive and significant if regional trends are included in the model. The size of the effect on verbal test scores ranges between 0.023 and 0.043 standard deviation 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) and (2).

The finding that the comprehensive school reform had 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 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, regional trends and

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

age-at-test dummies on both the mean and the variance, but only the effects of the comprehensive school reform are reported. The maximum-likelihood method produces very similar estimates for the effect of the reform on the mean scores as the linear regression model used in tables 2 and 3. The effects on means are significant only for the verbal test. The effects on the variance of the test scores are small but generally negative. In the math test the effect is close to zero. In the verbal and logical reasoning test the reform reduced the variance about 2.5 percent. None of these effects, however, are statistically significant.

[Table 4: EFFECTS ON MEAN AND VARIANCE]

As shown in tables 2 and 3, the estimated effect of the reform on test scores is somewhat sensitive to the inclusion of region-specific linear trends. This raises a concern that the standard differences-in-differences specification may fail to separate out pre-existing trends from dynamic policy responses, as pointed out by Wolfers (2006) in a different context. To explore this possibility, we re-estimated equation (1) by adding dummies for years two and three prior to the reform and separate dummies for each of the five years after the reform, thus omitting the dummy for one year prior to the reform. This flexible specification should trace out any pre-reform trends and dynamic policy responses following the reform. The results are reported in table 5 and a visual representation for the average test scores is plotted in figure 4. There is little indication of any pre-reform trends. However, especially the verbal test scores, and to a lesser extent the average test scores, increase significantly in the first year after the reform and stay at a higher level or even grow in the years following the reform. These results suggest that the pattern found in tables 2 and 3 does not reflect pre-existing trends but rather stems from the true reform effect.

[Table 5: THE EFFECT OF THE SCHOOL REFORM ALLOWING FOR LEADS & LAGS]

[Figure 4: LEADS AND LAGS]

6.2 Effects by parental background

Tables 6A and 6B examine the effect of the comprehensive school reform on average test scores by family background. Column (1) in table 6A estimates regression models similar to those reported in column (1) of table 3 but adds an indicator of parents' education and its interaction with the reform dummy. Parents are classified as being highly educated if at least

one parent 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. In column (2), we add linear region-specific trends to the regression. Since the interest here is in the difference of the effect of the reform between recruits from high- and low-education families, we can control for regional trends in a more flexible way in these regressions. In column (3), we add interactions of birth cohorts and regional dummies thus controlling also for any non-linear regional trends. After adding these interactions the main effect of the reform on test scores is no longer identified, but the difference of effect between recruits from high and low educated families still is. Finally, in column (4), we further introduce the full interactions of parental education with cohort and implementation region fixed effects as well as with the test age dummies.

According to table 6A, parental education has a clear effect on test scores. Men with highly educated parents score, on average, 0.275 standard deviations higher. The effect of the reform, now referring to the effect on men with less educated parents, is positive and larger than in table 2. However, the most remarkable result in table 6A is the negative and significant estimate for the interaction of parental education and comprehensive school reform. According to the results the reform had significantly less effect on men with more educated parents. This result is robust to including region-specific linear trends or full cohort region interactions although no longer significant once full interactions are introduced in column (4). The point estimates in table 6A indicate that the reform increased the test scores of men from low-education families by 0.031 – 0.047 standard deviation units and but little or no effect on men from high-education families. The results are qualitatively similar for individual tests (not reported in table) and again the effect is strongest in the case of verbal test scores where the reform increased the test scores of recruits from low educated families by 0.06 standard deviation units. This effect is sizeable, amounting to a quarter of the effect of parental education.

In table 6B we repeat the same analysis now using parents' income as the measure of family background. The parents' income is measured by summing the annual taxable income of both parents, inflating the incomes to the 2002 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 of the reform and parental earnings, but it makes it possible to

interpret the main effect of the reform in table 6B as the effect of the reform on sons from families with sample mean income. The results are qualitatively similar to those using parents' education. Men with richer parents tend to have higher average scores and the interaction between the parents' income and the reform dummy is negative in all models but column (4) where it is not statistically different from zero.

[Tables 6A and 6B: EFFECTS BY FAMILY BACKGROUND]

6.3 Magnitude of the effects

The results presented above suggest that the comprehensive school reform increased verbal test scores by approximately 0.04 standard deviation units. This may seem like a small increase and it is therefore instructive to put our results into perspective by comparing them to the effect sizes found with other educational reforms. The closest comparisons are the various studies on the effect of tracking. Duflo et al. (2008) report that tracking students within schools in Kenya led to an average increase of test scores by 0.14 standard deviation units. Comparing different countries, Hanushek and Wößmann (2006) find no effect of early school tracking on mean test scores.

Jacob and Ludwig (2008) survey the effect sizes of various educational policies targeted at children from low income families in the US. The effects of these policies on test scores vary from a high of 0.22, related to a large reduction in class size, to effects as low as 0.03 of policies such as Teach for America. In addition, the policies surveyed by Jacob and Ludwig (2008) typically target much younger children, where we should expect to see larger effects. Hence, in the light of these results, the effects of the Finnish comprehensive school reform are small but not out of line with results from other education policies, especially given that the effects here are measured on average four years after the completion of comprehensive school while other studies tend to focus on more immediate outcomes.

Perhaps the most interesting result reported above is the consistent positive effect of the reform on the test scores of recruits from low-education and low income families. As this implies a reduction in test score differences along socioeconomic lines, it is informative to compare the test score results to the earlier estimates of the effect of the comprehensive school reform on the intergenerational income elasticity.

Pekkarinen et al (2009) found that the Finnish comprehensive school reform decreased the elasticity of sons' earnings with respect to their fathers' earnings by as much as 0.066 log points. In this paper we report that the interaction of the parents' earnings and the school reform is -0.036 in a comparable regression equation explaining the standardized test scores (Table 6B, Column 3). This estimate can be scaled using earlier results on the effect of Army test scores on earnings. For example, Uusitalo (1999) calculated that a one standard deviation increase in these test scores implies about a six percent increase in earnings at age 33. Hanushek and Wößmann (2008) found that a one standard deviation increase in the International Adult Literacy test score increases earnings by about nine percent in Finland. Multiplying -0.036 with either of these estimates indicates that the effect of school reform on test scores only explains less than half a percent of the observed 6.6 percent decline in intergenerational income elasticity. Even after allowing for generous corrections for measurement errors in the test scores, it is obvious that the effect of the school reform on measurable cognitive skills is not sufficient to explain its effect on the earnings distribution.

7. 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.

Here, we estimate 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 difference-in-differences approach with 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 scores. Most interestingly, however, for all of these tests, the effect of the reform was positive

and significant in families where the parents had only basic education or low income, indicating a reduction in cognitive skills differences along socioeconomic lines.

We argue that the changes in the distribution of skills are likely to be due to the more academic curriculum content and the change in peer groups that especially affected the students from less-advantaged families. As is typical with any evaluation of real-life policy interventions of this scale, we cannot disentangle the relative importance of these factors. However, most realistic changes in school tracking policies involve changes in both the teaching content and peer groups almost by definition. Hence reliable estimates of the overall causal effect of school tracking policies are highly relevant both in terms of the empirical study of educational policies as well as the current policy debate on school tracking.

References

Aakvik, A., K. G. Salvanes, and K. Vaage (2010), "Measuring heterogeneity in the returns to education in Norway using an educational reform", *European Economic Review*, 54(4), 483-500.

Angrist, J., and J. S. Pischke (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton and Oxford.

Bertrand, M., E. Duflo, and S. Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119, 249–275.

Black, S., P. Devereux, and K. G. Salvanes (2010), "The more the smarter? Family size and IQ", *Journal of Human Resources*, 45(1), 33-58.

Brunello, G., and D. Checchi (2007), "Does school tracking affect equality of opportunity? New international evidence", *Economic Policy*, Oct 2007, 781-861.

Cameron, A., D. Miller, and J. B. Gelbach, (2008), "Bootstrapped-based improvements for inference with clustered errors", *Review of Economics and Statistics*, 90 (3), 414-427.

Duflo, E., P. Dupas, and M. Kremer (2011), "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya", *The American Economic Review*, 101 (5), 1739-1774.

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* 116, C63-C76.

Hanushek, E., and L. Wößmann (2008), "The role of cognitive skills in economic development", *Journal of Economic Literature*, 46 (3), 607-668.

Jacob, B., and J. Ludwig, (2008), "Improving educational outcomes for poor children", NBER Working Paper, 14550.

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* 316 (5832), 1717.

Lazear, E. (2001), "Educational Production", *Quarterly Journal of Economics* 116 (3), 777-803.

Lindqvist, E., and R. Vestman (2011), "The labor market returns to cognitive and noncognitive ability: Evidence from Swedish enlistment", *American Economic Journal: Applied Economics,* 3 (1), 101-128.

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* 95 (1), 414-424.

OECD (2003),*"Learning for Tomorrow's World: First Results from PISA 2003"*, OECD, Paris.

Pekkarinen, T., R. Uusitalo, and S. Pekkala (2009), "School tracking 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. Kieseppä, 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* , 162, 1904-1910.

Uusitalo, R. (1999), “Return to education in Finland”, *Labour Economics*, 6, 569-580.

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

Wolfers, J. (2006), “Did unilateral divorce laws raise divorce rates? A reconciliation and new results”, *The American Economic Review,* 96 (5), 1802-1820.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

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

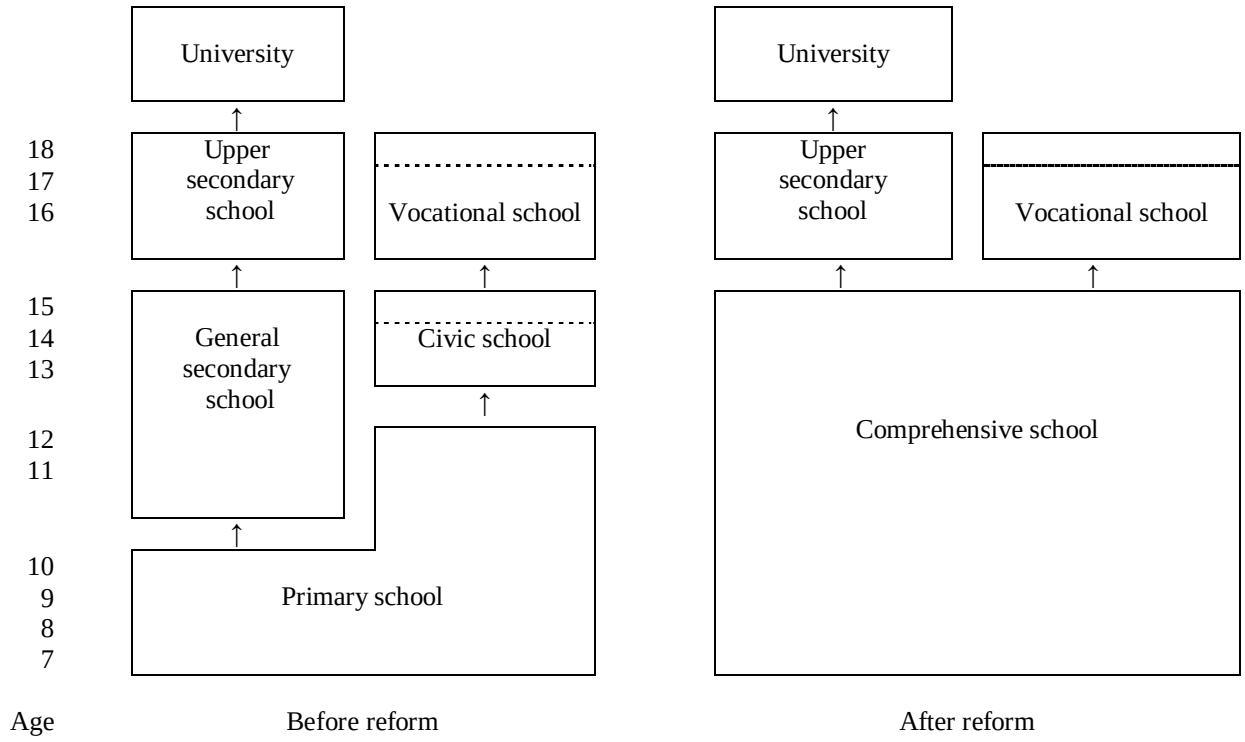


Figure 2 The implementation of the comprehensive school reform across regions 1972-1977

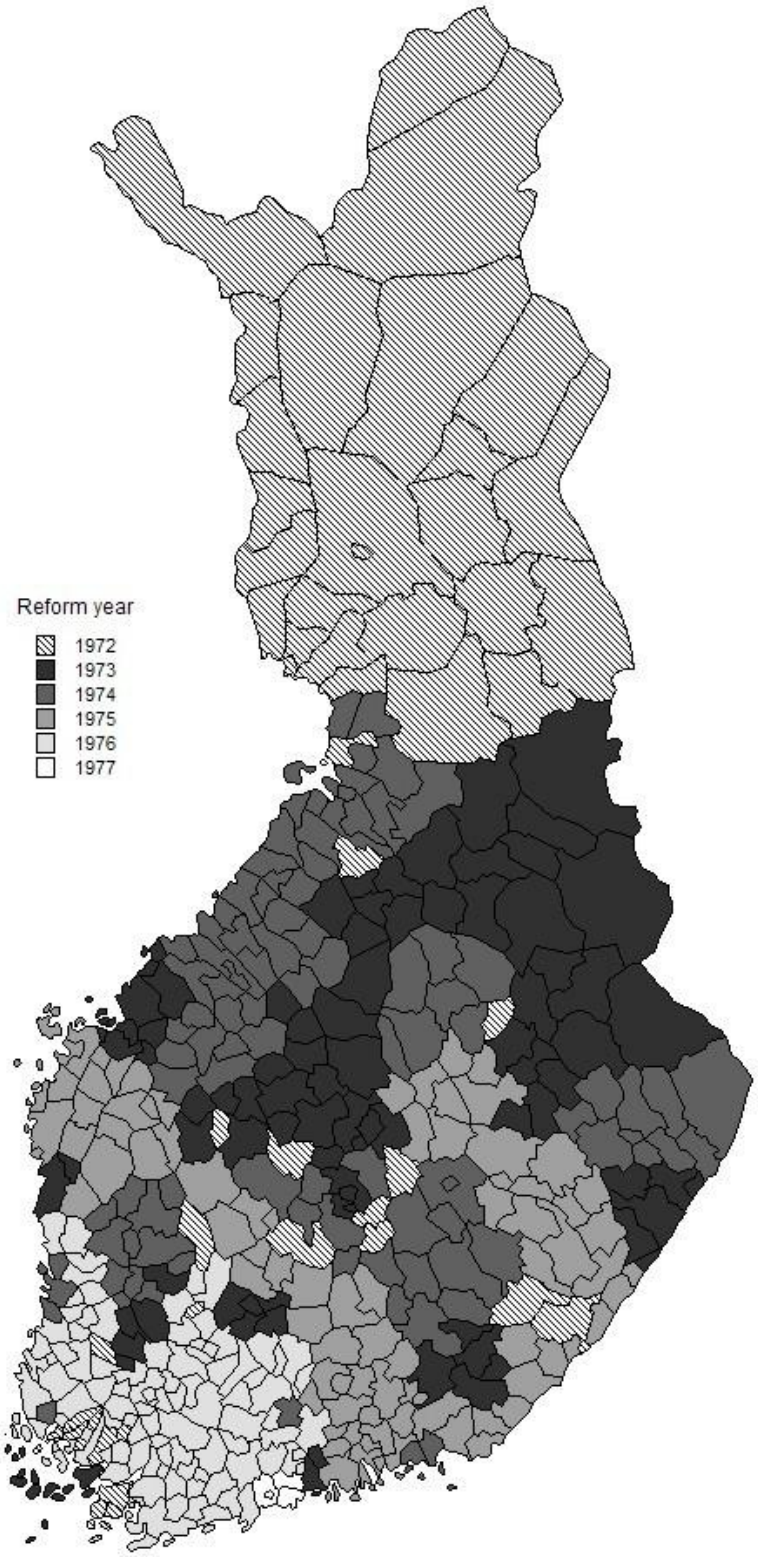


Figure 3 Distribution of the test scores

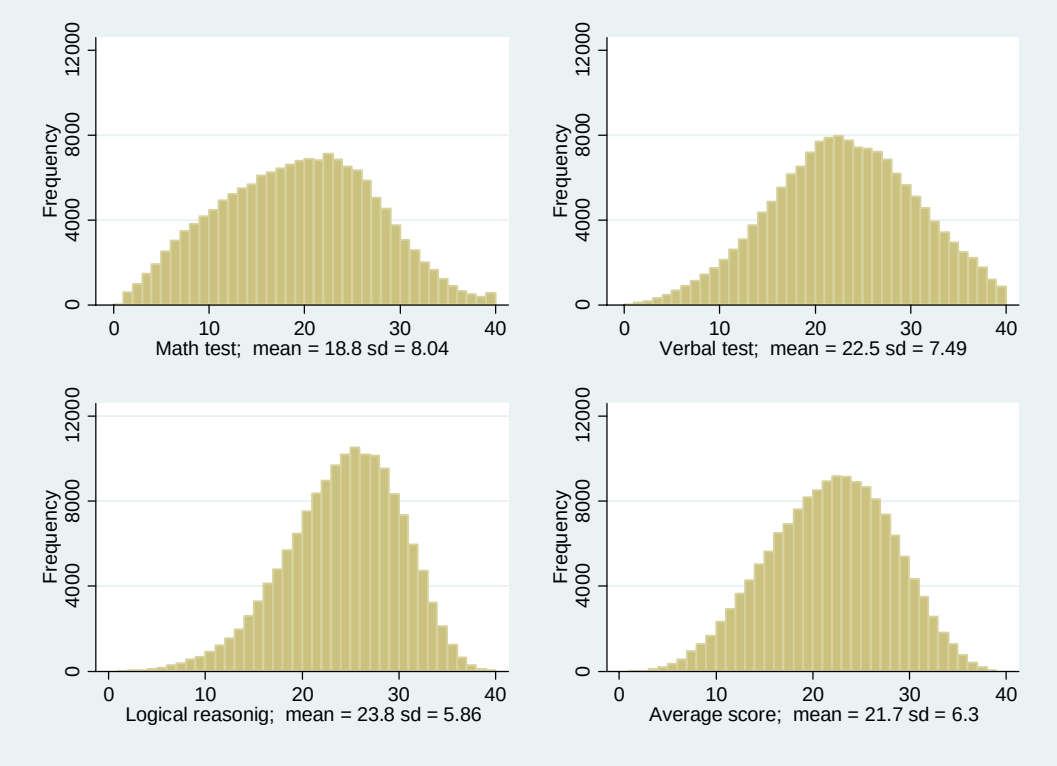
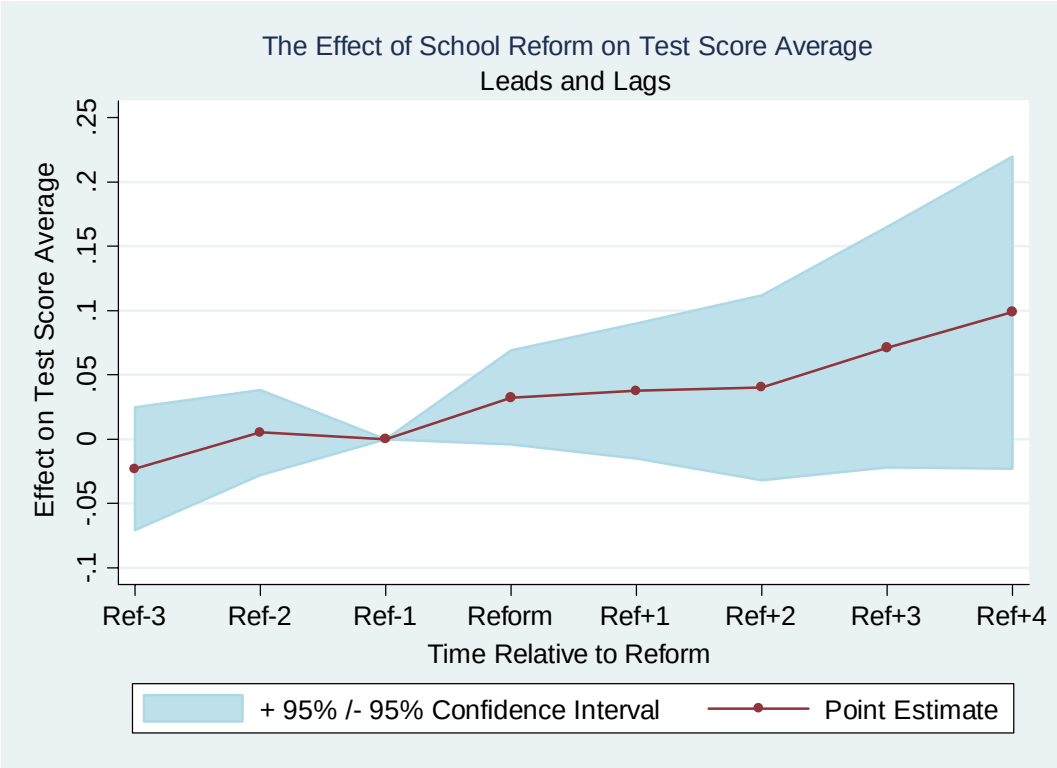


Figure 4 The effect of the school reform on the test score average, estimation around the reform date



Notes: The estimates in this graph represent a regression of the average test score on the cohort and implementation region dummies, age-at-test dummies, as well as separate dummies for the leads (up to 4 years after reform) and lags (up to 3 years before reform) of the reform implementation. The omitted category is "Reform-1". The plotted points are the estimates on the lead and lag dummies, and 95% confidence intervals are shown around the point estimates. Standard errors are clustered at the implementation region level and the critical values from a t-distribution with 4 degrees of freedom are used for inference.

Table 1 Standardized average test score by implementation region and birth cohort

	Region (Year when comprehensive school reform was implemented)						
Birth cohort	1972	1973	1974	1975	1976	1977	Total
1962	-0.21	-0.22	-0.16	-0.10	-0.01	0.19	-0.09
	[2,634]	[3,895]	[5,693]	[5,468]	[5,668]	[3,019]	[26,377]
1963	-0.10	-0.13	-0.07	-0.02	0.09	0.24	0.00
	[2,896]	[4,339]	[6,346]	[6,468]	[6,496]	[3,694]	[30,239]
1964	-0.11	-0.16	-0.06	-0.01	0.06	0.23	-0.01
	[2,865]	[4,299]	[6,238]	[6,483]	[6,723]	[3,977]	[30,585]
1965	-0.08	-0.11	-0.06	-0.00	0.11	0.22	0.02
	[2,715]	[4,036]	[5,995]	[6,304]	[6,290]	[3,889]	[29,229]
1966	-0.01	-0.04	0.01	0.03	0.14	0.29	0.07
	[2,185]	[3,314]	[5,117]	[5,579]	[5,550]	[3,590]	[25,335]
Total	-0.11	-0.14	-0.07	-0.02	0.08	0.24	0.00
	[13,295]	[19,883]	[29,389]	[30,302]	[30,727]	[18,169]	[141,765]
Parental education	0.48	0.46	0.50	0.51	0.59	0.69	0.53
Parental income	13,994	13,753	14,391	14,883	15,770	23,393	15,747

Notes: Table reports the average scores in the Finnish Army Basic Skills test data by region and cohort. Regions are defined by the year when the comprehensive reform took place in the region (see Figure 2). The unweighted average of the three subtest scores (math, verbal and logical reasoning) is standardized so that the sample mean is zero and the standard deviation is one. The number of observations in each cell is reported in square brackets, below the mean score. The shaded areas indicate cohorts that were affected by the post-reform educational system. The last two rows of the table report the share of parents with at least 12 years of education and average income of parents from the 1970, -75 and -80 census data inflated to 2002 euros using the consumer price index.

Table 2 The effect of the school reform on the test score average

	(1) Baseline	(2) Region & cohort	(3) Regional trends	(4) Family fixed effects
Reformed school	-0.095	0.010	0.025	0.023
	[-0.260,0.070]	[-0.017,0.038]	[-0.009,0.058]	[-0.023,0.069]
Birth year 1963		0.027	0.034	-0.009
		[0.001,0.053]	[0.003,0.064]	[-0.048,0.031]
Birth year 1964		0.030	0.043	-0.020
		[0.002,0.057]	[0.000,0.086]	[-0.074,0.033]
Birth year 1965		0.061	0.082	-0.026
		[0.031,0.092]	[0.024,0.141]	[-0.098,0.046]
Birth year 1966		0.078	0.108	-0.046
		[0.044,0.112]	[0.032,0.183]	[-0.138,0.047]
Reform region 1973		-0.034	-0.034	
		[-0.066,-0.001]	[-0.067,-0.001]	
Reform region 1974		0.043	0.045	
		[0.012,0.074]	[0.014,0.077]	
Reform region 1975		0.081	0.087	
		[0.049,0.114]	[0.053,0.120]	
Reform region 1976		0.193	0.201	
		[0.158,0.228]	[0.165,0.238]	
Reform region1977		0.343	0.355	
		[0.302,0.383]	[0.311,0.398]	
Lin. trend x region 1973			-0.002	-0.017
			[-0.026,0.022]	[-0.046,0.012]
Lin. trend x region 1974			-0.008	0.002
			[-0.031,0.015]	[-0.027,0.030]
Lin. trend x region1975			-0.019	-0.027
			[-0.043,0.006]	[-0.058,0.005]
Lin. trend x region 1976			-0.010	-0.023
			[-0.035,0.014]	[-0.054,0.009]
Lin. trend x region1977			-0.016	-0.038
			[-0.041,0.009]	[-0.071,-0.004]
Constant	4.557	4.343	4.315	3.904
R-squared	0.067	0.074	0.074	0.048
Observations	141,765	141,765	141,765	141,765

Notes: Sample includes birth cohorts 1962-66. The dependent variable is the unweighted average of the three subtest scores (math, verbal and logical reasoning) scaled into standard deviation units. All regressions include 13 age-at-test-dummies. Column 2 adds birth cohort and implementation region fixed effects. Column 3 further adds region-specific linear trends. Column 4 is estimated with cohort fixed effects, regional trends, age-at-test dummies and family fixed effects. Standard errors are clustered at the implementation region level and the critical values from a t-distribution with 4 degrees of freedom are used for inference. As non-standard critical values are used, the 95% confidence intervals are reported in parentheses below the point estimates.

Table 3 The effect of the school reform in different tests

	(1) Region & cohort	(2) Regional trends	(3) Family fixed effects
Math test	0.002 [-0.025,0.030]	0.015 [-0.019,0.048]	0.011 [-0.037,0.058]
Verbal test	0.023 [-0.004,0.051]	0.043 [0.009,0.077]	0.030 [-0.018,0.078]
Logical reasoning	0.006 [-0.022,0.033]	0.011 [-0.023,0.045]	0.027 [-0.025,0.078]

Notes: Sample includes birth cohorts 1962-66, n=141,765. Each cell of the table corresponds to a separate regression. Each subtest score is scaled into standard deviation units. The entries in the table represent the coefficients of a dummy variable indicating that the person attended the reformed comprehensive school. Column 1 includes cohort and implementation region fixed effects, and age-at-test dummies. Column 2 adds regional trends. Column 3 is estimated with cohort dummies, regional trends, age-at-test dummies and family fixed effects. Standard errors are clustered at the implementation region level and the critical values from a t-distribution with 4 degrees of freedom are used for inference. The 95% confidence intervals are reported in parentheses below the point estimates.

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

	(1) Math test	(2) Verbal test	(3) Logical reasoning test	(4) Average score in all 3 tests
OLS estimates				
Effect on the mean	0.015	0.043	0.011	0.025
	[-0.019,0.048]	[0.009,0.077]	[-0.023,0.045]	[-0.009,0.058]
ML estimates				
Effect on the mean	0.015	0.042	0.013	0.025
	[-0.019,0.049]	[0.008,0.076]	[-0.021,0.046]	[-0.009,0.058]
Effect on the log	-0.007	-0.025	-0.025	-0.024
Variance	[-0.056,0.043]	[-0.074,0.024]	[-0.074,0.025]	[-0.073,0.026]

Notes: Sample includes birth cohorts 1962-66, n=141,765. The entries in the table represent the coefficients of a dummy variable indicating that the person attended the reformed comprehensive school. Each regression model includes cohort and implementation region fixed effects, regional trends, and age-at-test dummies. The 95% confidence intervals are reported in parentheses below the point estimates

Table 5 The effect of the school reform in different tests allowing for leads and lags

	(1) Math test	(2) Verbal test	(3) Logical reasoning test	(4) Average score in all 3 tests
Reform - 3 years	-0.003	-0.032	-0.034	-0.023
	[-0.051,0.046]	[-0.080,0.017]	[-0.083,0.015]	[-0.071,0.025]
Reform - 2 years	0.006	0.005	-0.001	0.005
	[-0.027,0.040]	[-0.028,0.039]	[-0.035,0.032]	[-0.028,0.038]
Reform	0.016	0.055	0.020	0.032
	[-0.021,0.053]	[0.017,0.092]	[-0.017,0.058]	[-0.004,0.069]
Reform + 1 year	0.012	0.054	0.044	0.038
	[-0.041,0.064]	[0.000,0.107]	[-0.010,0.097]	[-0.015,0.090]
Reform + 2 years	0.003	0.070	0.045	0.040
	[-0.069,0.075]	[-0.002,0.143]	[-0.028,0.118]	[-0.032,0.112]
Reform + 3 years	0.024	0.098	0.083	0.071
	[-0.070,0.117]	[0.004,0.193]	[-0.012,0.177]	[-0.022,0.165]
Reform + 4 years	0.050	0.143	0.082	0.099
	[-0.072,0.171]	[0.020,0.265]	[-0.041,0.205]	[-0.023,0.220]
Constant	3.120	2.856	4.579	4.305
Observations	141,814	142,049	142,084	141,765

Notes: Sample includes birth cohorts 1962-66. Each column corresponds to a separate regression. Rows report the estimated coefficients of dummies for 3 and 2 years prior to the reform, as well as, for the immediate effect of the reform and 1, 2, 3, and 4 years after the reform. The omitted category is reform year – 1. Each regression includes cohort and implementation region fixed effects, and age-at-test dummies. Standard errors are clustered at the implementation region level and the critical values from a t-distribution with 4 degrees of freedom are used for inference. The 95% confidence intervals are reported in parentheses below the point estimates

Table 6A Effect of the school reform on average test score by parents' education

	(1) Cohort and region dummies	(2) Region specific linear trends	(3) Cohort x region interactions	(4) Full interactions with parental education
High educated parents	0.275	0.275	0.276	0.187
	[0.248,0.303]	[0.247,0.303]	[0.248,0.304]	[0.045,0.329]
Reform	0.031	0.047		
	[-0.003,0.065]	[0.007,0.086]		
Reform × high educated parents	-0.035	-0.035	-0.036	-0.031
	[-0.070,-0.001]	[-0.069,-0.000]	[-0.071,-0.002]	[-0.089,0.027]
Constant	2.270	2.237	2.920	2.962
Observations	126,977	126,977	126,977	126,977
R-squared	0.092	0.092	0.092	0.092

Table 6B Effect of the school reform on average test score by parents' income

	(1) Cohort and region dummies	(2) Region specific linear trends	(3) Cohort x region interactions	(4) Full interactions with parental income
Parents' income	0.325	0.324	0.327	0.246
	[0.299,0.352]	[0.298,0.351]	[0.300,0.354]	[0.109,0.384]
Reform	0.014	0.029		
	[-0.015,0.042]	[-0.006,0.064]		
Reform × parents' income	-0.034	-0.033	-0.036	0.002
	[-0.066,-0.002]	[-0.065,0.000]	[-0.069,-0.003]	[-0.053,0.058]
Constant	4.283	4.256	3.075	3.080
Observations	126,891	126,891	126,891	126,891
R-squared	0.101	0.101	0.102	0.102

Notes: Sample includes birth cohorts 1962-66. Each column corresponds to a separate regression. The dependent variable is the unweighted average of the three subtest scores (math, verbal and logical reasoning) scaled into standard deviation units. Parents' education is a dummy variable indicating that at least one parent had a degree higher than compulsory education. Parents' income is the log of average annual taxable income of parents from the 1970, -75 and -80 census data inflated to the 2002 price level using the consumer price index. Parents' income is measured as deviation from the mean log parents' income is used so the reform effect can be interpreted as the effect at the mean income level**.** Column 1 includes cohort and implementation region fixed effects, and age-at-test dummies. Column 2 adds linear regional trends. Column 3 adds dummies for the interactions between birth cohort and implementation region. Column 4 further adds the interactions of parental education (or income) with 1) birth cohort dummies, 2) implementation region dummies and 3) age-at-test dummies. Standard errors are clustered at the implementation region level and the critical values from a t-distribution with 4 degrees of freedom are used for inference. The 95% confidence intervals are reported in parentheses below the point estimates.

Appendix 1. Subsection scores

Table A1: Subsection scores in the Army test by cohort and region

Math score

	Reform year						
Birth cohort	1972	1973	1974	1975	1976	1977	Total
1962	-0.20	-0.19	-0.14	-0.09	0.01	0.15	-0.08
1963	-0.11	-0.10	-0.06	-0.00	0.08	0.22	0.01
1964	-0.12	-0.14	-0.06	-0.01	0.07	0.21	-0.01
1965	-0.08	-0.11	-0.06	0.00	0.09	0.21	0.01
1966	-0.01	-0.02	0.01	0.04	0.11	0.26	0.07
Total	-0.11	-0.12	-0.06	-0.01	0.07	0.21	0.00

Verbal score

	Reform year						
Birth cohort	1972	1973	1974	1975	1976	1977	Total
1962	-0.18	-0.19	-0.12	-0.10	-0.01	0.16	-0.08
1963	-0.06	-0.13	-0.04	-0.03	0.08	0.19	0.00
1964	-0.10	-0.14	-0.04	-0.00	0.05	0.19	-0.00
1965	-0.05	-0.10	-0.03	-0.01	0.11	0.16	0.02
1966	0.01	-0.04	0.01	0.02	0.11	0.24	0.06
Total	-0.08	-0.12	-0.04	-0.02	0.07	0.19	-0.00

Logical reasoning score

	Reform year						
Birth cohort	1972	1973	1974	1975	1976	1977	Total
1962	-0.18	-0.21	-0.16	-0.08	-0.02	0.20	-0.08
1963	-0.08	-0.12	-0.10	-0.01	0.06	0.24	-0.01
1964	-0.08	-0.15	-0.07	-0.02	0.06	0.23	-0.00
1965	-0.08	-0.09	-0.05	0.00	0.09	0.22	0.02
1966	-0.02	-0/04	0.01	0.03	0.15	0.30	0.08
Total	-0.09	-0.13	-0.08	-0.02	0.07	0.24	-0.00

Appendix 2. Issues related to the standard errors of the estimates

In the main parts of the paper we report standard errors based on the Moulton procedure and use for statistical inference a t-distribution with G-2 degrees of freedom, where G is the number of clusters (6 in our case). Clustering is done at the region, not at the region/cohort level. We also make a small sample adjustment suggested by Cameron, Miller and Gelbach (2008) inflating the residuals by $\sqrt{G/(G-1)}$. Since we use non-standard critical values we report condidence intervals instead of standard errors below the estimates.

After correcting the standard errors for clustering, the remaining concern is whether the cluster correction works given the small number of clusters and the potential serial correction of the errors. To assess the severity of the problem we followed the example of Bertrand, Duflo and Mullainathan (2004) and calculated mean residuals from equation 1 by cohort and region. We then estimated autocorrelations by an OLS regression of mean residuals on the lagged mean residuals. The resulting estimates for first order autocorrelations were low: .03 for the average score, .11 for math, -.09 for verbal and .07 for the logical reasoning test. These estimates may naturally be downward biased due to the short time series. However, testing the null of no autocorrelation by regressing the first-differenced residuals on their lagged values, as suggested by Wooldridge (2002), indicates no significant first order autocorrelation. In any case, even somewhat larger autocorrelation coefficients (e.g. around .2) only lead to modest over-rejection rates in the Monte Carlo simulations of Bertrand et al. (2004), Table 2.

For the sake of completeness we also calculated bootstrapped standard errors using the procedures described in Cameron et al. (2008). As the data are clustered by region, we used the block bootstrap and wild bootstrap that retain the cluster structure in the data and can be used with unequal cluster sizes. We did this bootstrapping the t-statistic, as recommended by Cameron et al (2008).

In the tables below, we compare the confidence intervals obtained from the various estimation procedures for the key parameters of our paper. We report the confidence intervals based on OLS, Stata cluster correction (with Stata default option that uses critical values from a t-distribution with G-1 degrees of freedom), and Moulton correction (with the small sample correction and critical values from a t-distribution with G-2 degrees of freedom). Finally, we also report bootstrapped estimates with 200 replications using the block bootstrap-t and the wild bootstrap-t.

Our conclusion from these experiments is that the OLS confidence intervals are clearly too narrow in the first column of Table A2 when no fixed effects are included. Other methods produce results that are qualitatively similar to each other, although the bootstrapped confidence intervals are somewhat wider. Once the fixed effects are included (Column 2), the confidence intervals produced by different

cluster correction procedures are quite similar, and even OLS confidence intervals appear to be quite accurate. Our interpretation is that this is mainly due to the low degree of serial correlation in the data. Confidence intervals based on the Moulton procedure appear to be the most conservative choice.

Table A2: Confidence intervals for Table 2, effect of the reform on average score

	Column 1 (b = -.095)		Column 2 (b = .010)		Column 3 (b = .025)	
Method	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI
OLS	-.106	-.085	-.008	.028	.003	.046
Cluster corrected	-.258	.068	-.008	.028	.010	.040
Moulton	-.260	.070	-.017	.038	-.009	.058
Block bootsrap-t	-.228	.136	-.017	.022	-.022	.089
Wild bootsrap-t	-.329	.138	-.007	.027	.011	.039

For the other tables the conclusions from our experimentation are very similar. Confidence intervals based on the OLS standard errors are typically only slightly narrower than the confidence intervals based on the parametric cluster corrections. Also, the wild bootstrap-t advocated by Cameron et al. (2008) produces confidence intervals that are close to the confidence intervals based on the usual cluster correction methods. The only method producing qualitatively different results is the block bootstrap-t that in some cases generates extremely wide confidence intervals. This finding resembles closely the Monte Carlo results of Cameron et al. (p.423), who note that the block bootstrap severely under-rejects the null when there are only a few clusters. Their explanation is that in some bootstrap replications only one treatment or control region is sampled, and for these replications the treatment dummy produces a perfect fit and zero residuals. The resulting Wald statistics in these re-samples are very large, which results in severe under-rejection rates. Our findings are similar. Wald estimates in some resamples are very large and hence the bootstrap confidence intervals very wide.

With the exception of the block bootstrap-t our experiments with bootstrapping indicate that the estimated confidence intervals are rather robust to the choice of cluster correction method. Again this is most likely due to the low degree of serial correlation in the data.

Table A3: Confidence intervals for Table 3, effect of the reform in the subtest scores

Without regional trends (Column 1)

	Math (b = .002)		Verbal (b = .023)		Logical (b =.006)	
Method	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI
OLS	-.016	.020	.006	.041	-.012	.024
Cluster	-.011	.015	-.002	.049	-.021	.032
Moulton	-.025	.030	-.004	.051	-.022	.033
Beta_block_t	-.020	.019	-.071	.043	-.041	.027
Beta_wild_t	-.010	.015	-.007	.052	-.018	.053

With regional trends (Column 2)

	Math (b = .015)		Verbal (b = .043)		Logical (b = .011)	
Method	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI
OLS	-0.007	.036	.021	.065	-.011	.033
Cluster	0.002	.027	.024	.062	-.017	.039
Moulton	-0.019	.048	.009	.077	-.023	.045
Beta_block_t	-0.021	.027	.000	.077	-.137	.079
Beta_wild_t	0.005	.024	.023	.064	-.028	.050

Table A4: Confidence intervals for Table 6A, interaction of reform and parents' education

	Column 1 (b = -.035)		Column 2 (b = -.035)		Column 3 (b = -.036)	
Method	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI
OLS	-.057	-.013	-.057	-.013	-.058	-.014
Cluster	-.072	.001	-.070	.000	-.073	.001
Moulton	-.070	-.001	-.070	.000	-.071	-.002
Beta_block_t	-.074	.097	-.068	.078	-.072	.094
Beta_wild_t	-.074	.004	-.071	.001	-.076	.004

Table A5: Confidence intervals for Table 6B, interaction of reform and parents' income

	Column 1 (b = -.034)		Column 2 (b = -.033)		Column 3 (b = -.036)	
Method	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI	Lower 95% CI	Upper 95% CI
OLS	-.054	-.013	-.053	-.012	-.057	-.015
Cluster	-.091	.024	-.090	.025	-.099	.026
Moulton	-.066	-.002	-.065	.000	-.069	-.003
Beta_block_t	-.081	.064	-.081	.062	-.093	.081
Beta_wild_t	-.097	.030	-.091	.025	-.103	.031

Appendix 3 Subsample results

The data used in the main parts of this paper include all conscripts who were born between 1962 and 1966 and who were found in the Army database. The army test scores are in the register from 1982 onwards. As it is possible to enter to military service as a volunteer before age 20, some men in the oldest cohorts served before the Army register was created. This generates some selectivity in the data.

In the final version of the paper we used the full sample and controlled for age at test. Controlling for age at test was done because age might have an independent effect on the test scores and because age of entering into the army is correlated with education. Volunteering into the army at age 19 is often convenient for high school graduates. Also education is one of the most common reasons for applying for deferment. In the data, the average test scores are higher for those who take the test before or after age 20 compared to those who take the test at age 20.

However, controlling for age at test is potentially a problematic solution if age at test is correlated with education and the comprehensive school reform affects the length of education, thereby making also age at test an endogeneous variable. Controlling for an endogenous variable or limiting the sample based on such variable would generally lead into biased results.

As a potential solution for the selectivity problem we restricted the data to men born between 1964 and 1966, i.e. those whose test scores we can observe even if they volunteered for early service at age 18 or 19. In order to balance the size of the treatment and the control groups in this limited sample the regions where the reform was implemented between 1972 and 1974 were also excluded from the analysis. The latter restriction also lessened the problem caused by unknown treatment status of persons who lived in different municipality in 1970 than in 1975. Such differences were relatively frequent because of migration and municipality mergers. The internal migration rates peaked in early 1970's and were about 25% higher in 1971-75 compared to 1976-80. In addition, altogether 71 of 518 municipalities that had existed in 1970 merged with their neighbouring municipalities between 1971 and 1975. Restructuring of local governments was much less intensive in late 1970s and hence the municipality codes that were used to determine the treatment status were more stable between 1975 and 1980.

In Tables A6, A7 and A8 we report the results from the restricted sample. The estimates are generally slightly larger than those presented in the main parts of the paper but qualitatively the restrictions made little difference. Eventually, these estimates were dropped from the version submitted for publication mainly because we calculated standard errors clustered by region and these restrictions would have resulted to even lower number of clusters.

Table A6: Reform effects using cohorts 1964-1966

VARIABLES	(1) Math	(2) Verbal	(3) Logical Reasoning	(4) Average score
Reform	0.0256	0.0788***	0.0264	0.0479*
	(0.0248)	(0.0247)	(0.0247)	(0.0247)
Constant	2.548***	3.238***	4.109***	3.434***
Observations	48,397	48,470	48,476	48,385

Notes: 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 region level.

Table A7: Effect of the reform on average test score by parents' education, cohorts 1964-1966

	(1) Region-specific linear trends	(2) Cohort x region interactions	(3) Full interactions with parental education
High ed. parents	0.315***	0.316***	0.0644
	(0.0178)	(0.0178)	(0.0717)
Reform	0.122***		
	(0.0286)		
Reform ×	-0.0713***	-0.0727***	0.0156
high ed. parents	(0.0209)	(0.0210)	(0.0353)
Constant	3.083	3.164***	3.104***
Observations	42,991	42,991	42,991

Table A8: Effect of the reform on average test score by parents' income, cohorts 1964-1966

	(1) Region-specific linear trends	(2) Cohort x region interactions	(3) Full interactions with parental income
Parental income	0.345***	0.347***	0.116*
	(0.0182)	(0.0183)	(0.0662)
Reform	0.0734***		
	(0.0255)		
Reform ×	-0.0366*	-0.0398*	0.0692*
parental income	(0.0212)	(0.0214)	(0.0354)
Constant	3.543***	3.249***	3.245***
Observations	3.543***	3.249***	3.245***

Notes: Parents' income is the 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.