The effects of delaying tracking in secondary school: evidence from the 1999 education reform in Poland

Maciej Jakubowski, Harry Anthony Patrinos, Emilio Ernesto Porta & Jerzy Wiśniewski

To cite this article: Maciej Jakubowski, Harry Anthony Patrinos, Emilio Ernesto Porta & Jerzy Wiśniewski (2016): The effects of delaying tracking in secondary school: evidence from the 1999 education reform in Poland, Education Economics, DOI: 10.1080/09645292.2016.1149548

To link to this article: http://dx.doi.org/10.1080/09645292.2016.1149548
The effects of delaying tracking in secondary school: evidence from the 1999 education reform in Poland

Maciej Jakubowski\textsuperscript{a}, Harry Anthony Patrinos\textsuperscript{b}, Emilio Ernesto Porta\textsuperscript{c} and Jerzy Wiśniewski\textsuperscript{d}

\textsuperscript{a}Faculty of Economic Sciences, Warsaw University, Warsaw, Poland; \textsuperscript{b}Education, World Bank, Washington, DC, USA; \textsuperscript{c}Analitica LLC, Miami, FL, USA; \textsuperscript{d}Ministry of National Education, Warsaw, Poland

ABSTRACT

Delaying tracking, extending students’ exposure to a general academic education and increasing their time on task on basic competences (reading, mathematics) could improve academic outcomes. To test the hypothesis that delayed vocational streaming improves academic outcomes, this paper analyzes Poland’s significant improvements in international achievement tests and the restructuring of the system which expanded general schooling. Estimates using propensity-score matching and difference-in-differences estimates show that delaying vocational education and increasing time on task have a positive and significant impact on student performance on the order of a standard deviation.

ARTICLE HISTORY

Received 14 December 2014
Accepted 27 January 2016

KEYWORDS

Tracking; time on task; student performance

JEL CLASSIFICATION

I20; I21; I24

1. Introduction

The early tracking of students by ability into different school types is hotly debated. Extending students’ exposure to a general academic education and increasing their time on task could improve academic outcomes. Many countries track pupils into academic and vocational schools at some point during their secondary schooling. Vocational training was an integral part of the education system in the planned economies of Central and Eastern Europe. The transition in many countries has led to changes in labor market demands. How curriculum tracking affects academic results, which are key to labor market success, is an issue that policy-makers have experimented with and researchers have tried to analyze.

In the 1950s to 1970s, many European countries reformed their education systems to delay or do away with selection of students into separate schools at an early age (early tracking) and the introduction of nationally unified curricula and comprehensive schooling. Examples include Finland, France, Norway, Sweden and the United Kingdom. The effects of such policy change have been difficult to evaluate as the changes have coincided with other major reforms or have been implemented simultaneously across the country.

Exploiting the successive implementation of a reform across municipalities in 1950s’ Sweden, which simultaneously replaced the academic and non-academic track with comprehensive schools, Meghir and Palme (2005) find that the reform increased schooling and earnings for students with low socio-economic background. In 1991, Sweden’s upper secondary school two-year vocational programs were transformed into three-year programs as a pilot before the reform was implemented all over the country four years later. This ‘natural experiment’ was evaluated in terms of years of upper

CONTACT Harry Anthony Patrinos\hpatrinos@worldbank.org

Supplemental data for this article can be accessed 10.1080/09645292.2016.1149548.
secondary education, university enrollment and the rate of inactivity. Results suggest positive effects on upper secondary education for those who lived in a pilot municipality in 1990 (Ekström 2002; Hall 2012).

Early tracking tends to benefit children with more educated parents (Bauer and Riphahn 2006). Experiments with delayed tracking have been shown to improve outcomes in Finland, Germany and Norway (Piopiunik 2013), particularly for the relatively disadvantaged (Aakvik, Salvanes, and Vaage 2010; Kerr, Pekkarinen, and Uusitalo 2012). Guyon, Maurin, and McNally (2012) evaluated a reform in Northern Ireland that led to a large increase in the share of students admitted to the elite track at age 11. They find a strong positive overall effect of this de-tracking reform on the number of students passing national examinations at later stages and a negative effect on student performance in non-elite schools that lost their most able students. Malamud and Pop-Eleches (2010, 2011) evaluate a reform in Romania which postponed tracking of students into vocational and academic schools by two years, finding no effect on university completion, labor market participation or earnings. For disadvantaged students, however, they find an increased probability to finish the academic track.

This paper contributes to the small literature that investigates the effect of timing of tracking on student performance by exploiting school reforms. In 1999, Poland reformed its basic education system in order to raise the level of education in society, increase educational opportunities and improve the quality of education. At that time, the new government restructured basic education by converting the eight-year primary school that was followed by vocational tracking into a six-year primary education followed by three years of lower general secondary education. Only after nine years of schooling would a decision be taken about what type of upper secondary education, academic or vocational, would follow. In other words, the new system postponed for one year the choice between general or vocational curriculum at the secondary level. Also, as part of the reform students dedicated more time for math and reading during the school week.

The purpose of our paper is to explain Poland’s significant improvement in international achievement tests (the Organization for Economic Co-operation and Development’s (OECD) Programme for International Student Assessment (PISA)) in the years 2000–2006. In math, Poland improved its score from 470 points in 2000, to 490 in 2003 and to 495 in 2006. Reading scores have steadily improved over time, from 479 to 508. In 2000 Poland ranked below the OECD reading average; in 2003 it reached the OECD average; and by 2006, Poland scored above average, ranking 9th among all countries in the world. During this period Poland was ranked 4th in terms of improvement in reading.

We use the variation created by the policy change in 1999 to test the impact on test scores over time. Specifically, we estimate a difference-in-differences (DD) model that compares the change in test scores of the likely vocational school students who were able to study in the general, academic track because of the change in school system. We find that, on average, the reform was associated with significant improvements. Poland improved its score in mathematics by 0.25 of a standard deviation, in reading, by 0.28 of a standard deviation, and in science, by 0.16 of a standard deviation. We confirm these results using our evaluation model – propensity-score matching and DD to create counterfactual scores for the group of likely vocational students in subsequent years – and the OECD’s PISA, an internationally comparable standardized student test conducted every three years to test reading, mathematics and science achievement of 15-year-olds. We decompose the results and estimate that two-thirds of the observed test-score differential between PISA 2000 and 2006 is associated with the changes in characteristics or the level or resources, while one-third reflects changes in the effect of characteristics and resources. At the school level, most is due to the change in hours of instruction. We use PISA data from 2000, 2003 and 2006, with 2000 as the baseline, since that wave of the study covered Polish students who studied within the old (pre-reform) system. We conclude that the reform is associated with an improvement in likely vocational students’ scores of about 100 points, or a whole standard deviation. We explore the implications using a 2006 special application of PISA in Poland that focused on 16- and 17-year-olds, and warn of the dangers of tracking (especially early tracking) to vocational education.
This paper is composed of eight sections: the second describes the policy change in Poland; this is followed by a description of test scores over time; next our hypotheses are presented; followed by a description of our empirical methods and data; the next section presents the average impact results; then we present additional analyses; we summarize our conclusions and discuss the policy implications in the final section.

2. Reform of 1998–1999

The government formed as a result of the parliamentary elections in 1997 set the following objectives for the education system: raising educational attainment in the society by increasing the number of those holding secondary and higher education qualifications; ensuring equal educational opportunities and supporting improvement in the quality of education viewed as an integral process of upbringing and instruction.

In order to achieve these objectives, a comprehensive reform of the education system was designed, which covered the structure of the school education system, the pupil assessment system, supervision and school governance. Also, the curricula were revised and schools started to dedicate more time for studying math and reading.

Before the reform (Figure 1), children completed one pre-school preparatory year and entered the eight-year primary school at age 7. Having completed primary education at the age of 15, young people could continue education in four-year general secondary schools, five-year technical secondary schools or three-year basic vocational schools. In order to select applicants, general and technical secondary schools conducted entrance examinations. Topics for compositions and math items were elaborated by the regional education authorities, which meant that exams conducted in different provinces differed from one another and thus were not comparable. There were no entrance exams in basic vocational schools. Entrance exams to the secondary schools could not serve as a tool for evaluation or monitoring the performance of school education in Poland. Tests did not provide for comparisons between different provinces or between consecutive years.

Students of secondary schools (both general and technical) were awarded a ‘maturity’ certificate upon passing the final exam at the end of secondary education. The maturity certificate gave them

![Figure 1](https://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx)

**Figure 1.** Structure of the Polish education system. Notes: Mature = ‘maturity’ final secondary school exam; ISCED = International Standard Classification of Education created by UNESCO to facilitate comparisons of education statistics and indicators across countries on the basis of uniform and internationally agreed definitions. Source: Authors.
access to higher education (though student applicants were also required to take entrance examinations). Basic vocational schools awarded only worker’s qualifications. In order to apply for admission to higher education programs, basic vocational school leavers were required to complete a two-year program in a supplementary technical upper secondary school and pass the maturity exam.

As a result of the reform, the eight-year primary school was replaced with a six-year primary school and, as a new type of school, a three-year lower secondary school (Figure 1). Thus, general education for all young people was extended by one year and the selection of applicants to various types of general and vocational upper secondary schools was, consequently, postponed by one year.

The change in the structure of the school education system and the related curricular reform were combined with the introduction of external examinations taken by all pupils at the end of primary and lower secondary education, and by secondary school leavers who aim to apply for admission to higher education. The latter take the (new) maturity exam which had replaced entrance exams in higher education institutions. That change in the mechanism of selection to tertiary education made it impossible to compare the data on applications and intake to universities before and after the introduction of the reform.

The assessment system focuses on monitoring math and reading and it has contributed to providing incentives for schools to dedicate more time for studying these two subjects during the school week. The age cohorts covered by PISA in 2000, 2003 and 2006 have been affected by the reform in different ways. The first group – those assessed in 2000 – was not affected by the reform. The group that was 15 years old in 2003 and was covered by the second cycle of PISA entered the primary school in the former system but attended the lower secondary school, the flagship of the reform. They did not take the final test in the sixth grade of the primary school. The test was administered for the first time in 2002, when they were already lower secondary students. The group covered by PISA 2006 was part of the reformed education system for most of their school careers. They took the final primary school test in 2003 and were prepared for the final lower secondary school exams a few weeks after PISA was administered in 2006.

The group covered by PISA 2000 consisted of first-grade students of the pre-reform secondary schools: general secondary schools, which students could enter only if they passed an entrance exam, vocational secondary schools and basic vocational schools, which were not highly regarded. The results of PISA 2000 in Poland showed a large variation in performance among schools, which was not surprising given that entry into secondary schools in the pre-reform system was determined by written entrance exams taken by primary school leavers. The groups covered by PISA 2003 and PISA 2006 consisted of students of the last (third) grade of the compulsory lower secondary school, and thus the results showed smaller variations among schools and larger ones among students within schools.

For most Polish students covered by the survey, PISA 2000 was the first experience in writing a test-item exam. Although PISA 2003 participants had not taken a test-item exam before, they had had some previous test experience in the form of mock exams that their teachers had introduced to prepare them for their upcoming final exams in the lower secondary school (Konarzewski 2004, 2008).

It has to be underlined that prior to the reform there were no national tests or exams that could be used for monitoring the performance and quality of education. PISA was the first international comparative study on student achievement in which Poland participated. The sustained increase in scores may be a sort of Flynn effect unrelated to the reform. But the increase in scores in Poland seems to be independent of developments in other central and eastern European transition countries.

Improvements in student performance in Poland, measured by PISA, have been impressive. In math, Poland improved its score from 470 points in 2000, to 490 in 2003 and to 495 in 2006. Reading scores have steadily improved over time, from 479 to 508. In fact, in the first assessment, Poland ranked below the OECD country average in reading. In 2003, Poland reached the OECD average; and by 2006, Poland scored above average, ranking 9th among all countries in the world. Science scores increased from 483 to 498.
Not all transition countries improved over time. Poland is the only country with consistent improvement over time. In fact, among the countries that participated in all three rounds of PISA, only Latvia and Poland improved. Latvia started at a lower level than did Poland, and its performance over time is impressive. However, while Latvia improved in reading between 2000 and 2003, its scores declined slightly between 2003 and 2006. Moreover, Poland is the only transition country to reach the top 10 countries in terms of performance in any subject in PISA. Latvia never reached the top 10 in any subject.

3. Empirical methods and data

We test whether the reform – specifically, the change in the structure of the school system, the delay of tracking and the extension of comprehensive general education – led to the improvement in test scores. Our main approach is based on propensity-score matching and reweighting. The propensity score reflects the probability of being assigned to one of the groups given a set of known characteristics. Rosenbaum and Rubin (1983) demonstrated that matching on the propensity score can balance distribution of the known characteristics across groups, so direct comparisons are more plausible.

We start with the assumption that one wants to compare survey results that are directly non-comparable because of differences in the distribution of observable characteristics. One can then calculate conditional expectations based on these characteristics and use them to calculate the difference of interest. However, when the number of distinct values of important covariates is high or when some of them are continuous, then any comparison of this kind becomes problematic: this is usually referred to as the ‘curse of dimensionality’. To resolve this problem, one can use propensity-score matching methods (Rosenbaum and Rubin 1983). Instead of matching multiple characteristics, the propensity score is balanced across comparison groups.

Originally, propensity-score matching methods were applied to solve selection problems, but in recent applications they were also used to adjust statistics across data sets (Tarozzi 2007). Similar methods were also applied earlier to compare whole outcome distributions before and after reweighting based on observable individual characteristics (DiNardo, Fortin, and Lemieux 1996). In this paper, when comparing whole distributions of student achievement, we use simple propensity-score weight adjustment. The counterfactual outcome distribution is obtained using kernel density estimators with weights given by formula (1):

\[ w = \frac{1 - \Pr (\text{Depvar} = 1)}{\Pr (\text{Depvar} = 1)}. \]

Tarozzi (2007) argues that such reweighting produces comparable outcome distributions. Depvar = 1 is defined as being a sample of interest, or ‘target’ sample, which, in this case, means the sample of PISA students in 2000. Depvar equals 0 for students sampled in 2003 or 2006, depending on a comparison made. Conditional probabilities are estimated using logit regression with a set of student and family characteristics defined in the same way in all waves of the PISA survey, and recoded to have similar categories. In addition, we considered sample weights that are important when one wants to make inferences about population effects. The PISA survey design was accounted for by multiplying propensity-score weights and survey weights.

3.1. Estimates of score change for students in different tracks

Reweighting produces factual and counterfactual distributions that are balanced in observable characteristics and can be compared across survey cycles. However, it is clear that the performance of Polish students could change for other reasons besides replacing the traditional secondary school tracks with lower secondary comprehensive schooling. The education reform of 1999/2000 modified
not only school structure but also curriculum and teacher compensation. But it has to be taken into account that change in the teachers’ salaries was introduced gradually and that it did not bring any noticeable effect in the structure of that professional group (age, qualifications, etc.). Also, the impact of the change of the curriculum should not be overestimated. There are differences in the prescribed number of lessons per week of, for example, Polish language and literature particularly between old basic vocational schools and new lower secondary schools. But PISA was always administered in early spring. Effectively, the difference was limited to half year education following different curricula. The difference in scores is too high to be explained by different curricula.

Our strategy is to assess how extending obligatory comprehensive education by one year affected the performance of students in different tracks. More specifically, we are interested in whether students who were in traditional vocational schools in 2000 would have similar scores in 2003 or 2006 in the newly established lower secondary comprehensive schools. That could be determined by matching vocational school students from 2000 with their counterparts in 2003 and 2006. In this way we can estimate the change in performance among students sharing characteristics common in each track. Then we look at the differential impact of the reform for students who were in different tracks in 2000. The change for vocational school students minus the change for general, or mixed vocational–general, school students could be attributed mainly to the introduction of lower secondary schools. That is, without the reform, 15-year-old students in vocational schools would not have had the opportunity to study in general programs. However, students in other tracks had this opportunity despite the reform. Students from general tracks can serve as a control group, and the difference in a simulated score change for them and for the former vocational school students could be attributed to postponing vocational education by one year.

Our approach to estimating the differential score change is similar to the DD method. This method compares outcome change in the group of interest (treatment group) with similar change in the control group. DD estimates of treatment effect take into account trends in the whole population that equally affect both groups. We calculate the difference between the achievement of students in vocational schools in 2000 and similar students in 2003 or 2006, and we subtract it from the difference between scores of secondary, general track students in 2000 and their counterparts in 2003 or 2006. Assuming that we are able to match similar students across waves of the PISA study, we can estimate how the reform affected students who, without the reform, would still be in vocational schools.

We use treatment-evaluation nomenclature (Lee 2005) to formally define the groups. The treatment is defined as a 15-year-old student in vocational secondary school in 2000. The control group is defined as 15-year-olds in general or mixed general–vocational secondary schools. We construct counterfactual groups of students from 2003 or 2006 samples based on their observable characteristics. A crucial assumption is that these observable characteristics constitute the main factors that explain differences in student achievement across treatment groups. This assumption is called ‘selection on observables’ in the literature. Bearing in mind that PISA collects a rich set of background characteristics that can often predict student performance, we believe that our assumption is well-founded and our approach is valid.

Let $Y_{it}$ be an outcome of an $i$-th individual in time $t = 0, 1$. We assume that some individuals were exposed to the treatment between $t = 0$ and $t = 1$, and write $D_{it} = 1$ if an $i$-th individual was exposed to the treatment. In the rest of this paper, we drop individual argument $i$ for simplicity. The DD model is formulated as:

$$\alpha = \{E(Y_1|D_1 = 1) - E(Y_0|D_1 = 1)\} - \{E(Y_1|D_1 = 0) - E(Y_0|D_1 = 0)\}. \quad (2)$$

A crucial assumption in this model is that a difference between transitory shocks in time $t = 0$ and $t = 1$ is mean independent of the treatment (Heckman, Ichimura, and Todd 1998; Abadie 2005). That means that without the treatment, the average outcome for the treated would change in the same way as the average outcome for the controls, or untreated. This assumption could be challenged if
groups differ in important characteristics. Thus, a conditional DD estimator is usually employed that controls for the set of covariates:

\[ \alpha_X = \{E(Y_1|X, D_1 = 1) - E(Y_0|X, D_1 = 1)\} - \{E(Y_1|X, D_1 = 0) - E(Y_0|X, D_1 = 0)\} \] (3)

The crucial assumption here is that quasi-experimental groups differ only by observable covariates. This condition eliminates any bias. Typically, the DD model is estimated using simple regression analysis, when any characteristic one wants to control for could be entered into the equation and made to interact with time and treatment (Gruber 1994; Meyer 1995). Another approach is to balance covariates across groups to make them more comparable, which can be achieved through matching methods (Rosenbaum and Rubin 1983; Heckman, Ichimura, and Todd 1998).

For our study, we need to find counterparts for the treatment and control groups in 2000 among students in lower secondary schools in 2003 or 2006. This can be achieved with matching methods where counterfactual \( t = 1 \) scores are constructed using scores of students with similar characteristics to those observed in \( t = 0 \). Usually, matching methods are used to make control and treatment groups more comparable, assuming that we have the same observations in each group in \( t = 0 \) and \( t = 1 \). In our case, we do not want to adjust for dissimilarities among treatment and control groups. We know that students who were in vocational schools differed from those in general schools, but we are interested in whether moving students from different tracks, who differ by assumption, into the one-type comprehensive lower secondary schools affected them similarly. Matching is used to adjust in time by drawing comparable groups from 2003 or 2006 samples, not for adjustments across quasi-experimental groups.

As already mentioned, when the dimension of \( X \) is high, then exact matching on covariates is not possible (the ‘curse of dimensionality’). In this case, individuals can be matched on one-dimensional propensity score \( P = P(D = 1|X) \), where \( D \) indicates treatment and \( P \) reflects the conditional probability of being treated (Rosenbaum and Rubin 1983). However, as we note above, we have to balance covariates not between treatment and control groups, which differ by assumption, but between waves of the survey. Only in 2000 were students treated, which means that they were separated into different types of secondary schools. After the reform, in PISA 2003 and PISA 2006, all students were in lower secondary comprehensive schools. Nevertheless, one can draw from 2003 and 2006 samples to find good matches and construct reference groups for students tested in 2000. We match using propensity score \( p^{2000} = P(T = 2000|X) \), reflecting the propensity to be in the PISA 2000 sample. Two propensity scores must be estimated: one measuring a propensity of being in a vocational school in 2000 for students tested in 2003 or 2006, and a second for being in a general (or mixed vocational–general) school in 2000 for students tested in 2003 or 2006. Thus, we have the propensity score for treated units (vocational school students) \( p^{2000}_T \) and the propensity score for controls \( p^{2000}_C \) (students in other tracks), both reflecting the propensity of being sampled in 2000 for students sampled in 2003 or 2006.

We define \( Y^1 \) as the score for students separated into tracks in secondary schools in 2000 and \( Y^0 \) as the score for students tested in 2003 or 2006. Now, the DD estimator could be defined by:

\[ \alpha_{DD} = \{E(Y^1|D = 1) - E(Y^0|p^{2000}_T, D = 1)\} - \{E(Y^1|D = 0) - E(Y^0|p^{2000}_C, D = 0)\}. \] (4)

In this equation, \( E(Y^1|p^{2000}_T, D = 1) \) and \( E(Y^1|p^{2000}_C, D = 0) \) are directly observed in the data, but \( E(Y^0|p^{2000}_T, D = 1) \) and \( E(Y^0|p^{2000}_C, D = 0) \) has to be constructed from 2003 or 2006 PISA samples using propensity scores. We first estimate the performance change for students in each type of secondary school in 2000 and their matched counterparts in 2003 or 2006. Then we compare these performance changes among students from different tracks. The difference between performance gains among students in the former vocational track and among students in other tracks is the DD estimator of the impact of abolishing the vocational curriculum for 15-year-olds. This estimator reflects the causal impact of the reform under the crucial assumption that the score change for students in the general track would be the same without the reform. This assumption is not directly testable,
however. For general track students, the curriculum did not change in a fundamental way, while other changes affected them as much as they did other students.

Propensity scores were estimated using logit regressions. To estimate the impact of the reform on student performance we applied 1-to-1 nearest neighbor propensity-score matching. This method matches to each treated observation one control observation with the closest value of the propensity score. A common support restriction was imposed, which means that if propensity-score distribution does not overlap at the bottom or top of the distribution, then observations with extreme propensity-score values will not be considered. This restriction only slightly affects the results in our case, but guarantees that proper matches were drawn from the 2003 and 2006 samples.

Finally, we need to decide which covariates to balance across surveys or use to draw counterparts of 2000 students in different tracks from 2003 and 2006 data. An obvious limitation is the availability of control variables that are identically defined across waves of PISA. Fortunately, PISA collects crucial variables reflecting students’ socio-economic background, including mothers’ and fathers’ education level and occupation, household cultural and educational resources, number of books at home and other household possessions. This information is summarized in the PISA index of economic, social and cultural status that we use to match students with similar socio-economic background. In addition, we also use student gender, age and the information about language spoken at home. Some of these indicators have a small number of missing observations. To ensure that the sample size and performance distribution are untouched by the matching exercise, missing values for matching covariates were imputed through the regression model with random component. This is similar to multiple imputation model with just one replication (Royston 2004). As the number of missing observations was small, imputation does not affect the final results but allows calculating performance differences for the whole sample of students.

The matching method assumes that all variables that are affecting the outcomes and that are also differently distributed across treated and control groups can be observed and included as matching covariates. Successful matching balances distribution of these variables across treated and controls groups and in this way eliminates the bias in outcome comparisons across groups. The appendix (see online supplemental file at http://dx.doi.org/10.1080/09645292.2016.1149548) provides detailed information and robustness analysis of our matching approach. In general, our method is able to reconstruct control samples without bias related to observed characteristics. However, matching approach does not account for selection on unobserved variables. Thus, if there are important factors that are unbalanced across treated and control groups but unobserved, the outcome comparisons will still be biased, even after matching.

Potential bias in matching estimates arising from selection on unobserved characteristics has to be addressed before concluding on the results. Although it is impossible to check unbalance on unobserved covariates as those are not available to a researcher, we are fortunate to have additional information that might be used to check how reliable the matching approach in our case is. Among the most important factors that affect selection of students into different schools are gender, socio-economic background, immigrant status and student achievement. We observe and control for the first three factors in our matching application. For the last factor, we do not have information on student achievement before selection to different tracks that would be available for all students. This is a serious threat to our approach as it is quite obvious that selection of students to upper secondary schools in Poland is mainly based on their achievement.

We use another study with similar data to estimate potential bias in our estimates due to selection on achievement. In Poland, the so-called national option in PISA 2006 extended the sample of students to 16- and 17-year-olds in different upper secondary school tracks. For these students we have additional information on their achievement from the national exams conducted before the selection to upper secondary schools. The exercise presented in detail in the online appendix assesses the potential bias in our matching estimates due to selection on achievement. In this analysis we use a similar approach as in the main part of the paper but it repeated it twice: with and without additional information on prior achievement. The results show that our main estimates...
can be biased downwards for general upper secondary schools, which only strengthens our con-
clusions, but crucial estimates of outcome improvements for vocational school students are not
biased. This is probably due to the fact that selection to vocational schools is not based on achieve-
ment but mainly on characteristics strongly correlated with socio-economic background that is con-
trolled for. Details of this analysis are discussed in the appendix (see online supplemental file at http://
dx.doi.org/10.1080/09645292.2016.1149548).

The PISA survey has a complex structure, similar to methods commonly used in other educational
surveys, with sampling conducted with different probabilities in two stages within separate strata. To
account for this, we use survey weights when calculating average outcomes for the treated students
in PISA 2000. In this manner the results are representative for the population of 15-year-olds in 2000.
To obtain correct standard errors that account for complex survey design and clustering of students
at the school level, the resampling method suggested in the survey documentation was used. PISA
provides replicate weights compatible with Fay’s adjusted balanced repeated replication (BRR). These
weights were constructed to reflect the sampling design, including any country-specific modifi-
cations, as well as non-response by students or schools (OECD 2002). In all calculations presented
in this paper standard errors were obtained by the suggested BRR method.

Results from the PISA study also have to take into account that students are answering randomly
assigned groups of test items (booklets), but responses are put into one common scale using psycho-
metric models. The performance of each student is reflected by five plausible values, which give
equally probable performance scores for individuals. Plausible values should not be used to judge
individual performance, but they provide unbiased estimates of achievement for whole populations
of interest. We follow the strategy suggested in the survey documentation of repeating each analysis
five times, with each plausible value used once to allow for measurement error in student perform-
ance. When using the multiple imputation method, we impute missing values once for each plausible
value and then repeat any estimation five times, once with each dataset containing one plausible
value and imputations obtained with this plausible value. That should guarantee that all imputation
errors, one in plausible values and the others in imputed covariates, will be taken into account (OECD

4. Results

Our analysis focuses on reading literacy, as performance in this domain is fully comparable across
PISA cycles. The results are presented for the whole sample and for the modal grade only, which
is the ninth grade in Poland. In PISA 2000, only the 9th grade was sampled; in PISA 2003 and
2006, students from the 7th, 8th and 10th grades were also sampled. The results suggest that stu-
dents in non-modal grades have a slight effect on the estimates. In the regression and matching
analysis, we simply adjust for student grade to account for these differences.

<table>
<thead>
<tr>
<th></th>
<th>Factual</th>
<th>Factual</th>
<th>Factual</th>
<th>Factual</th>
<th>Factual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>modal grade</td>
<td></td>
<td>modal grade</td>
<td></td>
<td>modal grade</td>
</tr>
<tr>
<td>Reweighting to 2000</td>
<td>2000</td>
<td>479.1</td>
<td>479.1</td>
<td>496.6</td>
<td>485.2</td>
</tr>
<tr>
<td>Mean score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change from 2000</td>
<td>–</td>
<td>–</td>
<td>17.5</td>
<td>6.1</td>
<td>22.8</td>
</tr>
<tr>
<td>Reweighting to 2000</td>
<td>2000</td>
<td>479.1</td>
<td>479.1</td>
<td>507.6</td>
<td>502.8</td>
</tr>
<tr>
<td>Mean score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change from 2000</td>
<td>–</td>
<td>–</td>
<td>28.5</td>
<td>23.7</td>
<td>34.4</td>
</tr>
<tr>
<td>Reweighting to 2003</td>
<td>2003</td>
<td>496.6</td>
<td>501.9</td>
<td>507.6</td>
<td>499.5</td>
</tr>
<tr>
<td>Mean score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change from 2003</td>
<td>–</td>
<td>–</td>
<td>11.0</td>
<td>2.9</td>
<td>11.6</td>
</tr>
</tbody>
</table>

Table 1. PISA results for Poland in Reading. Factual (with survey weights), reweighted to reference year (with survey and propensity-score weights), modal for modal grade.
Reweighting clearly lowers the mean scores of students in 2003 and 2006 (Table 1) while scores for students in the modal grade are slightly higher. When combined, these effects, which influence results in opposite ways, are positive, suggesting that overall student performance increased between 2000 and 2003 or 2006. For example, the change in factual scores (weighted only with survey weights) from 2000 to 2003 is 17.5, and from 2000 to 2006 is 28.5; but the change diminishes after reweighting to 6.1 and 23.7. However, after reweighting and taking students from the modal grade only, the gains are equal to 13.5 and 30.6. Thus, there is no doubt that increases in mean scores occurred from 2000 to 2003. The change between 2003 and 2006 is less clear. After reweighting, the initial difference of 11.0 (or 11.6 in modal grade) almost disappears. Nevertheless, we clearly observe substantial overall improvement after 2000.

While the change in mean scores is interesting, looking at the change in whole distributions gives a more detailed picture. Figures 2 and 3 show estimated factual distributions of scores in 2000, 2003 and 2006, together with reweighted scores for 2003 or 2006. The figures clearly show that the whole score distributions are ‘shifted’ to the right in 2003 and 2006 compared to 2000. This means that the difference in achievement across PISA cycles is not only among low achievers but also among high achievers. Poland thus closes the gap at all levels of performance. In PISA 2000, 24.5% of students scored in the top two reading proficiency levels, the fourth and fifth levels, compared to the OECD average of 31.8%. In 2006, this percentage increased to 34.7%, compared to the OECD average of 29.3%. Meanwhile, the percentage of Polish students below or at the first proficiency level was 23.3% in 2000, compared to the OECD average of 17.9%; and 16.2% in 2006, compared to the OECD average of 20.1% (OECD 2003, 2007).

What caused the ‘shift’ in the student score distribution? While extending compulsory comprehensive education can explain higher performance for low achievers, who were mostly in vocational tracks, explaining the improvement in performance among top achievers is more complicated. The questions are: did introducing lower secondary schools have an impact on students in former general secondary schools? And what was in the reform that resulted in such significant improvements in test scores?

### 4.1. Estimates of score change for students in different tracks

Results for DD propensity-score matching estimates of the effect of abolishing the tracking system for 15-year-olds in Poland are presented in Tables 2–4. Table 2 contains estimates of factual and counterfactual mean scores for all students in PISA 2000, 2003 and 2006. Results for students in vocational
and non-vocational tracks are also presented. Factual scores were weighted by survey weights provided in the official PISA data sets. Counterfactual scores were constructed using matching methods with survey weights taken into account, as described above. Not surprisingly, the counterfactual mean scores for all schools are similar to those reported earlier for the modal grade (Table 1). This shows that the choice between reweighting or different matching methods has no crucial impact on final estimates.

Results are summarized in Table 3, which shows the estimates of score improvement. These estimates assess trends in performance for all students and across groups of students who, without the reform, would be in different secondary tracks. Again, there is overall improvement of average performance among 15-year-olds in Poland. Score improvement for all students is large and significant, around 26 points from 2000 to 2006. Crucial estimates concern the hypothetical performance improvement from 2000 in different tracks. Performance improvement for potential students of former vocational schools (International Standard Classification of Education (ISCED) 3C) is simulated to be slightly below 100 points from 2000 to 2003 and 116 points from 2000 to 2006. This is more than one standard deviation of PISA scores in OECD countries, which is a dramatic improvement. Obviously, these estimates are statistically significant, supporting the hypothesis that 15-year-old students who, without the reform, would be placed in vocational tracks benefited greatly from the

![Figure 3](image). Change in reading literacy distribution: 2003–2006.

<table>
<thead>
<tr>
<th>Reading achievement</th>
<th>PISA 2000 factual weighted score</th>
<th>PISA 2003 factual weighted score</th>
<th>PISA 2003 matched counterfactual score</th>
<th>PISA 2006 factual weighted mean score</th>
<th>PISA 2006 matched counterfactual score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N)</td>
<td>(N)</td>
<td>(N)</td>
<td>(N)</td>
<td>(N)</td>
</tr>
<tr>
<td>All schools</td>
<td>479.1 (3654)</td>
<td>496.6 (4196)</td>
<td>483.1 (3650)</td>
<td>507.6 (5233)</td>
<td>504.8 (3653)</td>
</tr>
<tr>
<td>ISCED 3C vocational schools</td>
<td>357.6 (983)</td>
<td>–</td>
<td>453.3 (981)</td>
<td>–</td>
<td>473.5 (983)</td>
</tr>
<tr>
<td>ISCED 3B technical schools</td>
<td>478.4 (1491)</td>
<td>–</td>
<td>478.5 (1490)</td>
<td>–</td>
<td>498.2 (1490)</td>
</tr>
<tr>
<td>ISCED 3A general schools</td>
<td>543.4 (1180)</td>
<td>–</td>
<td>516.4 (1173)</td>
<td>–</td>
<td>532.0 (1179)</td>
</tr>
<tr>
<td>ISCED 3A and 3B schools</td>
<td>513.6 (2671)</td>
<td>–</td>
<td>498.3 (2668)</td>
<td>–</td>
<td>513.4 (2671)</td>
</tr>
</tbody>
</table>
reform. However, the benefits for students in other tracks are not that evident. Students in mixed-general schools (ISCED 3A and 3B) have similar scores in 2003 and improved by 20 score points in 2006. Students in the general track would potentially have lower scores in 2003 and similar performance in 2006. Note, however, that the results for score change for students in general schools (ISCED 3A) are probably biased downwards due to unobserved selection on achievement (see methods section and the online appendix). It is also worth noting that weighted sum of counterfactual results for all three upper secondary school tracks matches the overall results for all students in 2003 as well as in 2006. Thus, matching separately groups of students in 2003 or 2006 to students in different schools in 2000 gives, on average, the same result for the population which shows that matching does not introduce overall bias in the results.

These findings are in line with economic intuition. The short-term effects of the reform could be harmful for general school students who were mixed with low achievers in the newly introduced lower secondary schools. In the longer term, however, this negative impact disappears. It could be that segregation between and within lower secondary schools recreated the former stratification. It is clear that students in mixed-general schools benefited from the reform when one considers the general skills tested. The effects are again more evident over the long term, probably because of similar adjustments and mixing with high-achieving students. The positive effects among vocational school students were expected because, after the reform, these students spent more time learning non-vocational subjects. What is striking is the magnitude of the improvement at nearly one standard deviation and the speed with which students adapted to the new system. Clearly, adding just a few months of general education in the place of vocational education dramatically changes the general skills for a large number of students.

Relevant DD estimates of performance change for vocational school students are presented in Table 4. They are based on simple calculations from previous tables but clearly show the improvement of vocational school students relative to students in other tracks. The first row shows estimates of the relative performance change of vocational school students versus all students in other tracks. This is the most reliable comparison because it is based on the highest possible sample size. As noted above, the estimates show that the relative improvement in performance among vocational school students is higher than one standard deviation of international scores. Relative improvement in comparison to students in mixed general–vocational schools is slightly lower but still substantial.

These results are significant. Former basic vocational schools represented 22% of the students, while technical schools represented 36%. Thus, the reform that increased overall scores affected

---

**Table 3.** Propensity-score matching estimates of score change for students in different upper secondary school tracks.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>All schools</td>
<td>3.9 (5.2)</td>
<td>25.6 (5.1)</td>
</tr>
<tr>
<td>ISCED 3C schools (vocational schools)</td>
<td>95.6 (8.4)</td>
<td>115.9 (7.1)</td>
</tr>
<tr>
<td>ISCED 3B schools (technical schools)</td>
<td>−5.5 (7.8)</td>
<td>19.7 (7.5)</td>
</tr>
<tr>
<td>ISCED 3A schools (general schools)</td>
<td>−27.0 (7.6)</td>
<td>−11.4 (7.0)</td>
</tr>
<tr>
<td>ISCED 3A and 3B schools (general and technical schools)</td>
<td>−15.3 (5.4)</td>
<td>−0.2 (4.7)</td>
</tr>
</tbody>
</table>

**Table 4.** Relative score change (DD) for students in vocational schools.

<table>
<thead>
<tr>
<th>Relative score change</th>
<th>From PISA 2000 to PISA 2003</th>
<th>From PISA 2000 to PISA 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISCED 3C versus ISCED 3A+3B</td>
<td>110.9</td>
<td>116.1</td>
</tr>
<tr>
<td>ISCED 3C versus ISCED 3A</td>
<td>122.6</td>
<td>127.3</td>
</tr>
<tr>
<td>ISCED 3C versus ISCED 3B</td>
<td>101.1</td>
<td>96.2</td>
</tr>
</tbody>
</table>
positively and significantly more than a half of the school population, together with 58% of students attending former ISCED 3B and 3C schools.

There is thus no doubt that students who were in vocational tracks in 2000 would have scored much lower without the reform. The results show that the reform improved the overall mean performance of 15-year-olds in Poland, mainly by boosting the performance of students in former vocational tracks. Two questions remain for policy-makers: will the positive impact of the reform last, that is, will 15-year-old students in lower secondary schools still have higher achievement one or two years later, after they were again separated into tracks at the upper secondary school level? And what particular changes in curriculum or in the structure of the school system boosted student scores? These two issues are investigated below by using data from the PISA 2006 national option in Poland, which provides performance scores for 16- and 17-year-olds, and by employing decomposition analysis.

4.2. Additional analyses of performance among 16- and 17-year-olds

PISA offers an option to participating countries to conduct additional research using its framework and measurement tools. Poland opted to conduct this additional survey among 16- and 17-year-old students in 2006. After taking into account the difference in student age, the performance of 15-, 16- and 17-year-olds could be compared across educational tracks of upper secondary schools. This gives us a unique opportunity to compare achievement differences among 15-year-olds selected to different secondary schools before the reform and achievement differences among 16- and 17-year-olds who after the reform were also selected into different types of secondary schools but after one more year of comprehensive education.

Estimates of mean achievement by PISA cycle, grade and type of school program are presented in Table 5. First, 16-year-old students in the 10th grade score, on average, higher than do 15-year-olds in the 9th grade, and 17-year-olds in the 11th grade score higher than 16-year-olds. This is in line with intuition that older students perform better and the difference is around 7 to 8 points. However, when we look at the type of school program, it is clear that mainly students in general schools improved, while 11th grade students in vocational schools had even lower scores than those in the 10th grade. This seems to be counterintuitive, but there are two highly likely explanations. First, students change tracks, mostly in the 10th grade. Most of these students do not perform well at school and are forced to move to the vocational or mixed general–vocational track. Because of these changes, student achievement in mixed general–vocational or vocational upper secondary schools could be lower in the higher grades. Second, since students in vocational tracks devote more time to vocational training in higher grades, their general skills could decline. Consequently, slightly lower achievement in the vocational track is not that surprising.

<p>| Table 5. Mean achievement by PISA wave, grade and type of school program. |
|-----------------|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>Type of school program</th>
<th>2000</th>
<th>2003</th>
<th>2006</th>
<th>International 9th grade</th>
<th>National 10th grade</th>
<th>National 11th grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean achievement</td>
<td>479.1</td>
<td>501.9</td>
<td>513.5</td>
<td>520.1</td>
<td>528.3</td>
<td></td>
</tr>
<tr>
<td>Lower secondary school</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General secondary</td>
<td>543.4</td>
<td>513.5</td>
<td>580.8</td>
<td>592.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(% of all students)</td>
<td>42%</td>
<td>45%</td>
<td>11%</td>
<td>13%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General (profiled) secondary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(% of all students)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocational secondary</td>
<td>478.4</td>
<td>505.9</td>
<td>508.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(% of all students)</td>
<td>36%</td>
<td>29%</td>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocational basic</td>
<td>357.6</td>
<td>388.8</td>
<td>384.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(% of all students)</td>
<td>22%</td>
<td>16%</td>
<td>15%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
More important from the perspective of this paper is whether students in vocational schools perform better after the reform. In other words, we want to see if the positive effect of the reform for 15-year-olds remains substantial even after they are selected to different types of upper secondary schools one year later in the new system. Simple evidence presented in Table 5 suggests that the effect is positive but much smaller than demonstrated by our estimates for 15-year-olds. The simple difference in average performance for 15-year-old students in vocational schools in 2000 and 16-year-old vocational school students in 2006 is around 30 score points.

Note, however, that the share of population in vocational schools decreased from 22% in 2000 to 15% in 2006. Most of students who would probably go to vocational schools in 2000 attended in 2006 not only basic vocational but also general (profiled) secondary schools or vocational secondary schools. The achievement of students in these schools is much higher, above the performance level of students in vocational secondary schools in 2000. Thus, the performance improvement is substantial for 16/17-year-olds even after selection to different school types. It is at least above 30 points. On the other hand, even if the reform has sustainable benefits for students in vocational schools (although fewer students go to these schools compared to 2000), their performance level is still too low and the gap between these students and those in other types of upper secondary schools remains large. The evidence in Table 5 suggests also that students in these schools do not progress in reading literacy from grade 10th to 11th which poses questions about the effectiveness of these schools when it comes to teaching general skills.

4.3. Decomposition results

We estimate a simple education production function and decompose the changes over time in an attempt to show one of the ways the reform may have led to improved student achievement. Table 6 presents the results of production function estimates along with the decomposition results in reading. (Detailed decomposition results available upon request.) Overall, two-thirds of the observed test-score differential between PISA 2000 and 2006 is associated with the changes in characteristics or the level or resources, while one-third reflects changes in the effect of characteristics and resources. At the school level, most is due to change in hours of instruction. Generally, attending more than four hours of reading classes per week is associated with a higher score and this effect increased over time. In addition, there was a large increase in the proportion of students that received more than four hours of reading instruction, from 1% in 2000 to 76% in 2006.

The results are similar in the modified decomposition (Table 7). Most of the differential can be explained by school characteristics, particularly the increase in class hours for reading instruction that was part of the reform.

5. Conclusions

Extending the vocational training at the secondary school level has been advocated for many decades. The call for technical and vocational schooling used to be a standard recommendation promoted by international organizations and implemented by several countries. Unfortunately, the
enthusiasm for this approach was not based on any substantial evidence of its benefits to students. The important question is at which point of students’ educational career the choice of vocational and general track should be offered.

The 1999 Polish education reform program gave us the opportunity to assess the impact of the delay of tracking to vocational training on test scores. Nevertheless, the Polish reform adds years of required education and shifts the timing of tracking. This should be kept in mind when interpreting the results. The performance of Polish students before the reform, including those in vocational schools, was evaluated through the PISA 2000 study. The following PISA studies were testing students who after the reform were attending comprehensive general lower secondary schools. Our identification strategy was based on the fact that likely vocational graduates did not have that option in PISA 2003 and PISA 2006, which provided a comparison group for students in PISA 2000. Our empirical approach is based on the combination of propensity-score matching and DD estimation used to construct and compare potential outcomes of students before and after the reform.

Our results suggest that, on average, vocational schooling reduces test scores by a full standard deviation. While other aspects of the reform program no doubt helped improve Poland’s PISA scores (such as devoting more time to study math and reading), delayed entry into vocational education played a major role. We argue that the way to achieve better PISA scores is through delay of tracking and offering the same general program to all students.

We substantiated our findings by taking advantage of the application of PISA to 16- and 17-year-olds. We find that once vocational school options are available again, when students are 16, test scores decline for those students who enter the vocational track. Our evidence suggests that although these students still benefit from one additional year of comprehensive education, the results show that vocational schools do not further improve their general skills. While this goes a long way towards proving our initial findings, it also serves as a caution to policy-makers about the effectiveness of vocational schooling, particularly when it is not designed to improve math and reading skills. Those are skills that all students can learn, if given the opportunity; they are also the real vocational skills in the world of work today.

**Acknowledgements**

We benefited greatly from the comments of Mamta Murthi, Alberto Rodriguez, Joanna Sikora, Lars Sondergaard, seminar participants at the European Association of Labour Economists conference, PISA Research Conference, OECD and the World Bank, an anonymous referee and the editor. Views expressed here are those of the authors and should not be attributed to the authors’ respective organizations.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**ORCID**

Maciej Jakubowski © http://orcid.org/0000-0002-2512-3133

Harry Anthony Patrinos © http://orcid.org/0000-0002-4939-3568
References


