The Impact of the 1999 Education Reform in Poland[1]

Maciej Jakubowski
Organisation for Economic Co-operation and Development (OECD)

Harry Anthony Patrinos
World Bank

Emilio Ernesto Porta
World Bank

Jerzy Wiśniewski
Center for Social and Economic Research (CASE), Poland

March 2010

Abstract: Increasing the share of vocational secondary schooling has been a mainstay of development policy for decades, perhaps nowhere more so than in formerly socialist countries. The transition, however, led to significant restructuring of school systems, including a declining share of vocational students. Exposing more students to a general curriculum could improve academic abilities. This paper analyzes Poland's significant improvement in international achievement tests and the restructuring of the education system that expanded general schooling to test the hypothesis that delayed vocational streaming improves outcomes. Using propensity score matching and differences-in-differences estimates, the authors show that delayed vocationalization had a positive and significant impact on student performance on the order of one standard deviation.

JEL Classification Codes: I21, I28
Keywords: education, streaming, tracking, curriculum

* We have benefited greatly from discussions with and comments from Mamta Murthi, Alberto Rodriguez, Joanna Sikora, Lars Sondergaard, and participants at seminars at the European Association of Labour Economists conference, PISA Research Conference and the World Bank. The views expressed here are those of the authors and should not be attributed to the World Bank Group. Address all correspondence to Harry Anthony Patrinos at hpatrinos@worldbank.org.

1. Introduction

The vocationalization of the curriculum is a major feature of education plans since the post war period. It is argued that vocational skills are needed for job creation and productivity. The common sense view is that vocational education is necessary for a country to modernize and acquire the technical skills needed for economic development (Psacharopoulos 1997). Many countries have developed vocational education systems. Socialist countries integrated vocational schooling into the overall economic planning system, assigning them to different ministries. The upshot was that employment was guaranteed in those models. After the transition began, however, the link between vocational education and employment was broken, leaving vocational students lacking in job opportunities and skills demanded by the workforce.

Indeed, vocationalization has been under attack for many decades. The social costs of vocational education may not match the social benefits associated with it (Psacharopoulos 1987). The argument that vocational education would bring industrialization and jobs was challenged as the "vocational school fallacy" (Foster 1965). More importantly, the "vocational skills" needed in the world of work today are not the old traditional skills linked to specific jobs; rather, they are the critical thinking and "learning to learn" skills (see Murnane et al. 1995), exemplified by success in math, reading and science, for example.

A small empirical literature does suggest that that there are advantages of targeted vocational *training* programs (Karlan and Valdivia 2006). Evaluations from the randomized programs in USA show modest effects at best (Heckman et al. 1999). While impacts in developing countries are more positive, most of those programs are not experimental (Betcherman et al. 2004). A subsidized program in the Dominican Republic showed no impact on employment, though a marginally significant impact on hourly wages and on the probability of health insurance coverage, conditional on employment (Card et al. 2007). A randomized training program for disadvantaged youth in Colombia raises earnings and employment for both men and women, with larger effects for women (Attanasio et al. 2009).

Fewer evaluations have been undertaken on the impacts of vocational education. Earlier assessments of vocational education programs in a number of countries have shown that most graduates of such schools go to university rather than entering manual occupations (Psacharopoulos and Loxley 1985). The expansion of vocational education in Sweden in the 1990s led to positive effects on years of upper secondary education but no significant effect on the probability to continue to higher education (Ekström 2002).

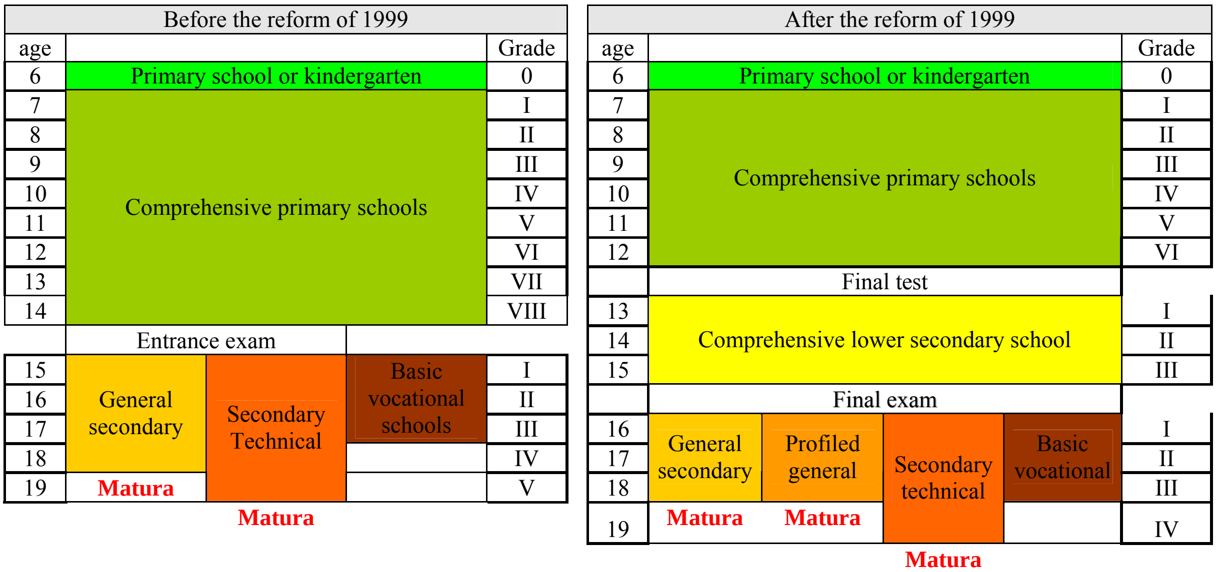
Poland is a good case for an evaluation of learning outcomes associated with vocational secondary schooling. Between 2000 and 2006 Poland improved in PISA test scores by 0.16 to 0.28 of a standard deviation in science, math and reading. This was preceded in 1999 with the reform of the education system. The old 8-year primary school that was followed by early vocational tracking, was converted into a 6-year primary education followed by three years of lower general secondary education. Only after 9 years of schooling would a decision about what type of upper secondary education (academic or vocational) be made. In other words, the new system postponed the choice of type of curriculum at the secondary level for one year. 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 difference model that compares the change in test scores of the likely vocational school students that were able to study in the general, academic track because of the change in school policy. We find, using propensity score matching and differences in differences that, on average, the reform was associated with significant improvements. We explore threats to identification using among other things decomposition analysis. 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 results for 16 and 17 year-olds, and warn of the dangers of early vocationalization.

The organization of the paper is as follows. Section 2 describes the policy change in Poland. Section 3 describes the increase in test scores over time. Our hypotheses are presented in Section 4. Section 5 describes our empirical methods and data. Section 6 presents the average impact results. Additional analyses are presented in section 7. Finally, we summarize our conclusions and discuss the implications in section 8.

2. Reform of 1998-1999

The Ministry of Education (1998) presented the goals of the reform as: (1) to raise the level of education; (2) to ensure equal opportunities; and (3) to support improvements in quality. The reform covered the following: restructuring the system; changes in administration and supervision; introduction of core curricula; independent assessment and examination; school finance; and qualification requirements for teachers. The structural reform led to the introduction of the lower secondary "gymnasium" as a new type of school. The previous structure, comprising eight-year primary followed by four-year secondary or three-year vocational school would be replaced with a system described as 6+3+3 (Figure 1). This meant the duration of primary school was reduced to 6 years. Following primary, pupils would continue in a three-year gymnasium and only upon completion would they move on to three-year secondary or a two-year vocational. The structural reform postponed the choice of the direction of education at the secondary level for one year. With the clear division of the education system into stages, pupil achievement could be assessed reliably through tests and examinations.

Figure 1: Structure of the Polish Education System



The age cohorts covered by PISA in 2000, 2003 and 2006 have been affected by the introduction of the reform in different ways. The first group (2000) was not affected by the reform. The group that was 15 years of age in 2003 and was covered by the second cycle of PISA started their education in primary school in the former system but attended the gymnasium, which was part of the new structure. They did not take the final test in the sixth grade of primary school. The test was for the first time administered in 2002, when they were already gymnasium students. Finally, the group covered by PISA 2006 has attended reformed schools. They took the primary school final test in 2003 and were prepared for the final gymnasium exams a few weeks after PISA was administered in 2006.

The group covered by PISA 2000 consisted of the first grade students of the pre-reform secondary schools: general lyceum (one had to pass an entrance exam to enter), secondary vocational school and basic vocational school (with very low prestige). The results of PISA 2000 in Poland showed huge variations among schools, which was not surprising at all, as the pre-reform system was based on a strong selection of primary school leavers. However, the group covered by PISA 2003 (and PISA 2006) consisted of students of the last (third) grade of compulsory gymnasium. Not surprisingly, the results showed smaller variations among schools.

3. Relative Increase in Scores

Improvements in student outcomes 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 (Table 1). Reading has seen a steady improvement over time, from 479 to 508. In the first assessment Poland ranked below the OECD country average in reading. In 2003, Poland reached the OECD average. Moreover, by 2006, the latest assessment, Poland scores above average, ranking 9th among all countries in the world. In science, the scores are 483, 498 and 498.

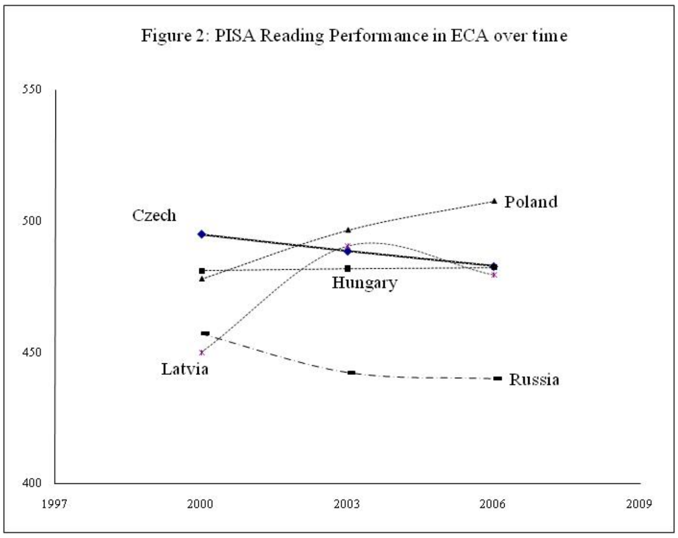
Table 1: Top 10 Reading over Time, PISA

	2000		2003		2006	
1	Finland	549	Finland	543	Korea	556
2	Netherlands	537	Korea	534	Finland	547
3	Canada	535	Canada	528	Hong Kong	536
4	Hong Kong	532	Australia	525	Canada	527
5	Australia	528	Liechtenstein	525	New Zealand	521
6	Ireland	528	New Zealand	522	Ireland	517
7	New Zealand	526	Ireland	515	Australia	513
8	Japan	525	Sweden	514	Liechtenstein	510
9	United Kingdom	524	Netherlands	513	Poland	508
10	Korea	522	Hong Kong	510	Sweden	507

4. Hypotheses for Explaining Change over Time

Several factors could explain these changes. The claim of this paper is that many statistics produced from PISA are not fully comparable across years if specific empirical questions are to be answered. If one is interested in international comparisons to assess the effectiveness of countries' educational policies then only samples equivalent in a distribution of important student and family characteristics could be compared. To give an example, if countries differ in a distribution of parental education which strongly affects student outcomes, then it is not valid to compare mean performance in these two countries to conclude whether one has more effective education policy than the other. Most likely the difference in mean performance depends more heavily on the difference in parental education than on the policy itself. Thus, any comparison on unadjusted samples could be policy irrelevant or unhelpful. Similarly, if one wants to compare achievement levels in a particular country in different years, then one needs to adjust samples to make them fully comparable. While PISA organizers try to maintain sampling schemes that are the same in all countries and years, it is difficult to preserve similar samples across time, especially when the school system changes.

Not all transition countries improved over time. Among participating countries in Eastern Europe, Poland is one of the best performers, with a solid improvement over time. Poland is the only country with consistent improvement over time (Figure 2). In fact, among the five that participated in all three rounds of PISA, only Latvia and Poland improved over time. Latvia started at a lower point than did Poland and its performance over time is impressive. However, unlike Poland, in reading Latvia improved between 2000 and 2003, but slightly declined between 2003 and 2006.



Reform led to improvement – through delay of vocational, more education-relevant inputs (hours, motivation, better teachers, etc.). This paper concentrates on Poland and tries to recognize differences between the country's samples in 2000, 2003 and 2006. Unadjusted score distribution is compared across years. After this purely descriptive analysis adjustments to make samples comparable are proposed and semi-parametric methods employed to produce equivalent score distributions. Finally, the differences-in-differences method is applied to test whether extension of comprehensive education was the main reason for score improvement over time.

Students are more accustomed to taking tests and teachers are preparing students for tests. Rigorous academic testing was not the norm prior to the 1999 reforms. Soon after the reforms tests became more important and regular. This exposure to assessments may have prepared students, thus making them better "test takers."

5. Empirical Methods and Data

We test whether the reform – more specifically the change in the structure of the school system – led to the improvement in test scores, through the delay of vocational education. Our main approach is based on propensity score matching and reweighting. Assume that one wants to compare survey results that are directly non-comparable because of differences in the distribution of observable characteristics. Then one can 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 some of them are continuous, then any comparisons of this kind become problematic. This is known as the "curse of dimensionality." To resolve this problem propensity score matching methods were proposed by Rosenbaum and Rubin (1983). They showed that instead of using multiple covariates one can use the *propensity score* that reflects the probability of being sampled to one of the groups conditional on covariates. Originally, propensity score matching methods were applied to solve selection problems but in recent applications they were also used to adjust statistics across datasets (see Tarozzi 2007). Similar methods were also applied earlier to compare whole outcome distributions before and after reweighting based on observable individual characteristics (DiNardo, Fort 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:

$$w = \frac{1 - \Pr(\text{Depvar} = 1)}{\Pr(\text{Depvar} = 1)}$$

Tarozzi (2007) argues that such reweighting produces comparable outcome distributions. *Depvar*=1 is defined as being in a sample of interest ("target" sample), which in our 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 (recoded to have similar categories). Additionally, we considered sample weights that are of importance when one wants to make inferences about population effects. PISA survey design was accounted for by multiplying propensity score weights and survey weights.

As covariates we used gender, age, mothers and father's education, the highest value of the International Socio-Economic Index among parents, number of books at home, and grade. Usually, researchers also control for immigrant status; however, there is a negligible number of migrants in the Polish sample. Missing data were imputed using the multiple imputation approach (Royston 2004). Results without any imputation were qualitatively similar, while less precise because of smaller sample sizes.

Estimates of score change for students in different tracks

Reweighting produces factual and counterfactual distributions, which 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 the introduction of comprehensive schooling. The education reform of 1999/2000 modified not only school structure but also curriculum, teacher compensation and many other things. Thus, the test score change cannot be solely attributed to the replacement of old-type tracks in secondary schools by lower secondary schools for 15-year-olds.

To deal with that, our strategy is to assess how the extension of 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 old-type vocational schools in 2000 would have similar scores in 2003 or 2006 in newly established lower secondary comprehensive schools. That could be investigated by matching vocational school students from 2000 with their counterparts in 2003 and 2006. This way an estimate of performance change for students sharing characteristics common in each track can be obtained. Having that, 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. The point is that without the reform 15-year-old students in vocational schools would not have 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 schools students could be attributed to the postponing of vocational education by one year.

Our approach to estimate the differential score change is similar to the differences-in-differences (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. In our case, 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.

To define it formally, we use treatment evaluation nomenclature (see Lee 2005). The treatment is defined as being a 15-year-old student in vocational secondary school (*szkoła zawodowa*) in 2000. The control group is defined as 15-year-olds in general (*liceum ogólnokształcące*) or mixed general-vocational (*technikum*) secondary schools. We have to 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 explaining differences in student achievement across treatment groups. This assumption is called "selection on observables" in the econometric evaluation literature. Having in mind that PISA collects a rich set of background characteristics, which are strong predictors of student performance, we believe that this assumption is fulfilled 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 the paper we drop individual argument i for simplicity. The differences-in-differences model is given by:

$$\alpha = \{E(Y_1 \mid D_1 = 1) - E(Y_0 \mid D_1 = 1)\} - \{E(Y_1 \mid D_1 = 0) - E(Y_0 \mid D_1 = 0)\}$$

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 (see Abadie 2005; Heckman, Ichimura and Todd 1998). It means that without the treatment the average outcome for the treated would experience the same change as the average outcome for the controls (or untreated). That assumption could be demanding if groups differ in important characteristics. Thus, usually conditional differences-in-differences estimator is employed which controls for the set of covariates:

$$\alpha_X = \{E(Y_1 \mid X, D_1 = 1) - E(Y_0 \mid X, D_1 = 1)\} - \{E(Y_1 \mid X, D_1 = 0) - E(Y_0 \mid X, D_1 = 0)\}$$

The crucial assumption here is that quasi-experimental groups differ only by observable covariates and this condition eliminates any bias caused by this. Typically, the differences-in-differences model is estimated using simple regression analysis when any characteristic one wants to control for could be entered into the equation and interacted with time and treatment (Meyer 1995; Gruber 1994). 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).

In our case, 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 whether moving students from different tracks, who differ by assumption, into the one-type comprehensive lower secondary schools affected them similarly. Matching in our case is used to adjust in time by drawing comparable groups from 2003 or 2006 samples, not for adjustments across quasi-experimental groups.

When 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 again indicates treatment and P reflects the conditional probability of being treated (see Rosenbaum and Rubin 1983). However, as we noted 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 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 need to be estimated. One measuring a propensity of being in a vocational school in 2000 for students tested in 2003 or 2006, and the other 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_T^{2000} and the propensity score for controls P_C^{2000} (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 of 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:

$$\alpha_{DD} = \left\{E(Y^1 \mid D=1) - E(Y^0 \mid P_T^{2000}, D=1)\right\} - \left\{E(Y^1 \mid D=0) - E(Y^0 \mid P_C^{2000}, D=0)\right\}$$

In this equation $E(Y^1 | D=1)$ and $E(Y^1 | D=0)$ are directly observed in the data, but $E(Y^0 | P_T^{2000}, D=1)$ and $E(Y^0 | P_C^{2000}, D=0)$ have to be constructed from 2003 or 2006 PISA samples using propensity scores. We first estimate the performance change for students in each type of school in 2000 and their matched counterparts in 2003 or 2006. Then we compare these performance changes across students from different tracks. The difference between performance gains of students in the former vocational track and for students in other tracks is the differences-in-differences estimator of the impact of abolishing 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. Two kinds of propensity score matching were then employed: nearest neighbor matching 1-to-1 matching and kernel matching. The first method matches to each treated observation one control observation with the closest value of the propensity score. The kernel method constructs values for matched counterparts by weighting control observations by their proximity in the propensity score to the treated observation, using a kernel function (we used Epanechnikov kernel with bandwidth 0.6; see Becker and Ichino 2002 for details of the Stata procedure used). In both methods 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 rarely 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 here is availability of control variables which are identically defined across waves of PISA. Fortunately, PISA collects crucial variables reflecting student socio-economic background, including the HISEI index (highest of mother or father international socio-economic index), mother and father ISCED education level, and number of books at home; in addition, student gender, age, grade they attended at the time of PISA survey, and family structure, were also used as covariates. Some of these indicators, mainly HISEI index, parental education levels, and family structure, have a small number of missing observations. To keep the sample size and performance distribution untouched by the matching exercise, missing values for matching covariates were imputed through multiple imputation models (Royston 2004).

We use Poland's data from the OECD's Program for International Student Assessment (PISA), an internationally comparable standardized student test conducted every three years to test reading, mathematics and science achievement of 15-year-olds. PISA assesses and compares student achievement based on a standardized and highly reliable framework. It assesses the knowledge and skills essential for full participation in society and the world of work. It is conducted near the end of compulsory education.

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. This complexity should be taken into account by using probability weights when calculating point estimates and by adjusting for clustering and strata design when estimating standard errors. However, there is little advice in the literature on how to account for survey design in matching methods (Zanutto 2006). In our case, we use survey weights when calculating average outcomes for the treated students in PISA 2000. This way the results are representative for the population of 15-year-olds in 2000. Also, students answer a randomly assigned groups of test items (booklets), but responses are put into one common scale using psychometric models. The performance of each student is reflected in five plausible values which give equally probable performance scores. Plausible values should not be used to judge individual performance but provide unbiased estimates of achievement for whole populations of interest. Each analysis should be repeated five times with each plausible value used once to take into consideration measurement error in student performance. We follow this strategy, which should guarantee that all imputation errors, one in plausible values and the others in imputed covariates, will be taken into account (OECD 2002). The final set of variables used in this analysis are re-sampling replicate weights used in the calculation of standard errors. Intra-cluster correlation violates an assumption needed for the unbiasedness of the analytical method of calculating standard errors based on the variation of the sample. Re-sampling methods such as bootstrapping, Jackknifed Repeated Replication and Balanced Repeated Replication serve as alternative means of calculating standard errors. These methods calculate sampling variance by re-sampling the same sample to mimic re-sampling of the original population. Replicate weights are alternative sample weights which represent a sub-sample based on the original sampling design. PISA provides replicate weights compatible with Fay's adjusted Balanced Repeated Replication (BRR), constructed to reflect the sampling design including any country specific modifications and non-response by students or schools (OECD 2002). Standard errors were obtained by the BRR method.

Decompose Change over Time

A simple decomposition analysis is undertaken in order to attempt to explain one of the possible pathways be which the reform may have led to improved student achievement. We decompose reading scores over time PISA 2000 and 2006 to explain to what extent the increase in scores is due to changes in characteristics and what proportion is due to changes in returns to characteristics. A simple education production function is estimated, which relates various inputs affecting student learning to measured outputs. We specify and estimate education production functions that relate students' achievement to individual, family and school inputs. We then proceed to decompose the over time test score gap into an explained component (accounting for student, family, and school characteristics) and an "unexplained" – or returns, or the efficiency by which the country is able to convert characteristics into student learning outcomes as measured by test scores – component (Oaxaca 1973; Blinder 1973).

6. Results

In what follows, we focus on reading literacy as only performance in this domain is fully comparable across PISA cycles. Performance in mathematics can be compared across 2003 and 2006 only because the 2000 assessment framework was later modified. Science performance in 2006 cannot be related to previous cycles as the framework was completely changed in 2006. The results are presented for the whole sample and for the modal grade only, which is the 9th grade in Poland. In PISA 2000, only the 9th grade was sampled while in PISA 2003 and 2006 students from 7th, 8th and 10th grade were also sampled. The results suggest that students in non-modal grades slightly affect the estimates and should be taken into account. In the regression and matching analysis we simply adjust for student grade to take into account these differences.

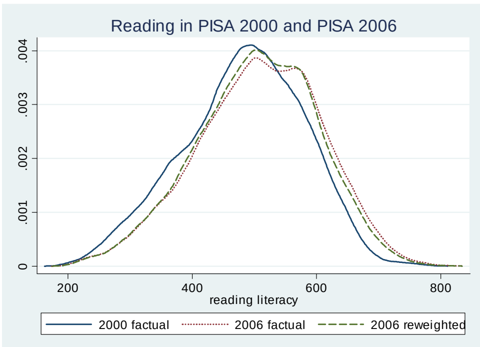
Reweighting clearly lowers the mean scores of students in 2003 and 2006 (Table 2). On the other hand, scores for students in the modal grade are slightly higher. These effects, which influence results in opposite ways, when combined are positive, suggesting that overall student performance increased 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 it is 28.5; but it diminishes after reweighting to 6.1 and 23.7, respectively. However, after reweighting and taking students from the modal grade only, the gains are equal to 13.5 and 30.6, respectively. Thus, there is no doubt that increases in mean scores occurred in Poland 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.

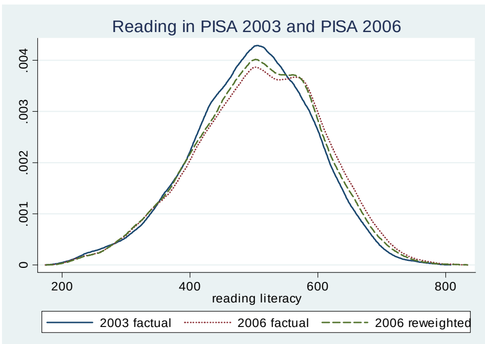
Table 2: PISA 2000, 2003, 2006 results, Poland, Reading Factual (survey weights), reweighted to the reference year (with survey and propensity score weights), and modal for modal grade

	Factual	*Factual Modal grade*	*Factual*	*Reweighted*	*Factual Modal grade*	*Reweighted Modal grade*
Reweighting to 2000	*2000*		*2003*			
Mean score	479.1	479.1	496.6	485.2	501.9	492.6
Change from 2000	-	-	17.5	6.1	22.8	13.5
Reweighting to 2000	*2000*		*2006*			
Mean score	479.1	479.1	507.6	502.8	513.5	509.7
Change from 2000	-	-	28.5	23.7	34.4	30.6
Reweighting to 2003	*2003*		*2006*			
Mean score	496.6	501.9	507.6	499.5	513.5	506.9
Change from 2003	-	-	11.0	2.9	11.6	5.0

The change in mean scores is obviously interesting but looking at the change in whole distributions gives a more detailed picture. The whole score distributions are “shifted” to the right in 2003 or 2006 compared to 2000 (Figure 3). This means that the difference in achievement across PISA cycles is not only for low-achievers but also for high-achievers. This way Poland closes the gap at all levels of performance. While in PISA 2000 the percentage of students in the top two reading proficiency levels (4th and 5th) was 24.5 percent (compared to OECD average of 31.8 percent), this number in Poland increased in 2006 to 34.7 percent (compared to OECD average of 29.3 Percent). At the bottom of the distribution, the percentage of Polish students at the 1st or below proficiency level was 23.3 percent in 2000 (17.9 percent in OECD) and 16.2 percent in 2006 (20.1 percent in OECD) (OECD 2003, 2007). That poses interesting questions about what caused the "shift" of the student score distribution. While extension of compulsory comprehensive education can explain higher performance for low-achievers who were mostly in vocational tracks, it is more complicated to explain higher performance of top-achievers. The question is whether the introduction of lower secondary schools could have an impact on students in former general secondary schools and what was in the reform that increased scores so significantly.

Figure 3: Change in Reading Distribution





<u>Estimates of score change for students in different tracks</u>

Results for differences-in-differences propensity score matching estimates of the effect of abolishing tracking system for 15-year-olds are presented in Tables 3, 4 and 5. Table 3 contains estimates of factual and counterfactual mean scores for all students in PISA 2000, 2003 and 2006. In addition, results for students in vocational and non-vocational tracks are presented. Factual scores were weighted by survey weights provided in the official PISA datasets. Counterfactual scores were constructed using matching methods with survey weights taken into account in the manner described above.

Not surprisingly, the counterfactual mean scores for all schools are similar to those reported earlier (see Table 2). Moreover, results for kernel matching and 1-to-1 matching are also similar. They differ slightly because of different matching methods and various number of matched control observations (provided in parentheses) but give qualitatively similar conclusions. This shows that the choice between reweighting or different matching methods has no crucial impact on final estimates.

Results are summarized in Table 4, where the estimates of score improvement are presented.[2] These estimates assess trends in performance for all students and across groups of students who without the reform would be in different secondary tracks. We see again overall improvement of average performance among 15-year-olds in Poland. Score improvement for all students is remarkable, at 16 to 18 points from 2000 to 2003 and around 35 to 36 points from 2000 to 2006. Crucial estimates are for the hypothetical performance improvement of 2000 in different tracks. Performance improvement for potential students of former vocational schools is

[2] The numbers presented in the third row, after the name of the comparison and matching method, are showing how these differences were calculated from the results presented in Table 3. In each case, the difference was calculated by taking a counterfactual performance score of matched student from the 2003 or 2006 samples and subtracting from it the factual score of students tested in 2000. Standard errors for these differences were calculated by employing the BRR method, which properly accounts for complex survey design (stratification, clustering, and response adjustments).

simulated to be higher than 100 points from 2000 to 2003 and 120 points from 2000 to 2006. This is more than one standard deviation of PISA scores in OECD countries, which is a dramatic improvement, hardly comparable to effects of any known educational policy. Obviously, these estimates are statistically significant, supporting the hypothesis that 15-year-old students who without the reform would be placed in vocational tracks heavily benefited from the reform. However, the benefits for students in other tracks are not that visible. Students in mixed-general schools improved their scores only slightly in 2003 and noticeably in 2006 only. Students in the general track would potentially have lower scores in 2003 and similar performance in 2006.

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 newly introduced lower secondary schools. In the longer term, this negative impact disappears. It could be that teachers adjusted their methods to more diversified classrooms or that segregation between and within lower secondary schools recreated the former stratification. Students in mixed-general schools obviously benefited from the reform when one considers general skills tested in PISA. Effects are again more visible in the longer term probably because of similar adjustments and mixing with high-achieving students. Positive effects for vocational school students were expected because after the reform they spend much more time learning non-vocational subjects. However, what is interesting is the enormous magnitude of the effect that is about one standard deviation of PISA international scores. Again, it is not surprising because of the change in classroom time allocated after the reform to the general instead of vocational subjects. However, what is surprising is that counterparts of vocational school students "adapted" so quickly to the new system. In other words, it is striking that just a few additional months of comprehensive instead of vocational education changes general skills for an important number of students so dramatically.

Table 3: Factual and counterfactual scores of students in different upper secondary tracks

Reading achievement	PISA 2000 factual weighted mean score (no of obs)	PISA 2003 factual weighted mean score (no of obs)	PISA 2003 matched counterfactual score (no of matched obs)		PISA 2006 factual weighted mean score (no of obs)	PISA 2006 matched counterfactual score (no of matched obs)	
			Kernel matching	*1-1 matching*		*Kernel matching*	*1-1 matching*
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
All schools	479.1 (3654)	496.6 (4196)	497.9 (4151)	495.2 (2528)	507.6 (5233)	514.9 (5229)	514.1 (3056)
ISCED 3C schools	357.6 (983)	-	466.7 (4010)	460.5 (926)	-	484.3 (5141)	474.4 (1090)
ISCED 3B schools	478.4 (1491)	-	491.4 (4150)	487.7 (1527)	-	507.3 (5163)	501.8 (1823)
ISCED 3A schools	543.4 (1180)	-	525.6 (4064)	524.9 (1233)	-	543.0 (5221)	547.0 (1376)
ISCED 3A and 3B schools	513.6 (2671)	-	507.3 (4157)	507.0 (2206)	-	524.8 (5233)	520.5 (2609)

Note: Standard errors are given in parentheses and were obtained from bootstrapping (kernel matching) or analytically (1-1 matching). * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Finally, we turn to relative improvement or differences-in-differences estimator of performance change for vocational school students. Relevant estimates are presented in Table 5. They are based on simple calculations from the tables above but clearly show the improvement of vocational school students versus score change for 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 of the highest possible sample size. As we noted above, the estimates show that the relative improvement in performance of vocational school students is higher than one standard deviation of international scores (100). Relative improvement in comparison to students in mixed general-vocational schools is slightly lower but still substantial.

Table 4: Propensity score matching estimates of score change for students in different upper secondary school tracks

Reading achievement	Score change: PISA 2003 – PISA 2000		Score change: PISA 2006 – PISA 2000	
	Kernel matching	*1-to-1matching*	*Kernel matching*	*1-to-1 matching*
	(1) - (3)	*(1) - (4)*	*(1) - (6)*	*(1) - (7)*
All schools	18.8 (4.3)	16.1 (4.5)	35.8 (4.4)	35.0 (4.5)
ISCED 3C schools	109.2 (5.8)	103.0 (5.8)	126.8 (5.7)	116.9 (6.3)
ISCED 3B schools	13.0 (5.7)	9.3 (6.5)	28.9 (5.8)	23.4 (7.2)
ISCED 3A schools	-17.8 (5.4)	-18.5 (4.3)	-0.4 (5.1)	3.6 (5.0)
ISCED 3A and 3B schools	-6.3 (4.3)	-6.6 (4.3)	11.2 (4.2)	6.9 (4.4)

Notes: Propensity score matching with common support restriction; Standard errors are given in parentheses and were obtained through BRR method accounting for complex survey design.

Summing up, there is no doubt that students who were in vocational tracks in 2000 would score 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 and mixed general-vocational tracks. There are two remaining question which are of policy relevance. One is whether this large positive impact of the reform was long lasting. More precisely, the question is whether 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. Secondly, it seems interesting to find out what particular changes in curriculum or organization of the school system boosted student scores. These two issues are investigated below by utilizing data from the PISA 2006 national option in Poland which provides performance scores for 16- and 17-year-olds and by employing decomposition analysis.

Table 5: Relative score change (differences-in-differences) for students in vocational schools

Relative score change	from PISA 2000 to PISA 2003		from PISA 2000 to PISA 2006	
	Kernel matching	*1-1 matching*	*Kernel matching*	*1-1 matching*
ISCED 3C versus ISCED 3A+3B	115.5	109.6	115.7	110.0
ISCED 3C versus ISCED 3A	127.1	121.5	127.2	113.3
ISCED 3C versus ISCED 3B	96.2	93.7	98.0	93.5

Additional Analyses

PISA gives an option to participating countries to conduct additional research using its framework and measurement tools. Poland utilized this option in 2006 to conduct surveys among 16- and 17-year-old students (see Federowicz 2007). These students were mostly in upper secondary school, which opens up a possibility for comparisons with students in lower secondary schools (from the international sample). After taking into account the difference in student age, one can compare the performance of 15-, 16- and 17-year-olds, also across educational tracks of upper secondary schools. In other words, knowing the achievement at the end of lower secondary schools we can compare how much students updated their skills during the time of education in different types of upper secondary schools.

The estimates show, 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-old in the 11th grade score higher than 16-year-olds (Table 6). This is in line with intuition that older students are more able to solve PISA tests. However, when we look at the type of school programs, it is clear that mainly students in ISCED 3A schools improved, whereas 17-year-old students in vocational schools have even lower scores. This seems to be counterintuitive but there are two possible, highly likely explanations. Firstly it has to be noted that students change tracks, mostly in the 10th grade, and these are mostly low performing students who are forced to move to the vocational or mixed general-vocational track. Because of such changes in the population, student achievement in mixed general-vocational or vocational upper secondary schools could be lower in higher grades. Second, students in ISCED 3C tracks devote more time for vocational training in higher grades. Therefore, their general skills tested in PISA could be diminished. Consequently, slightly lower achievement in ISCED 3C is not that surprising.

Box plots presented below summarizes score distribution for the categories presented in Table 6 (Figure 5). This time data for vocational upper-secondary schools and general profiled (mixed) upper-secondary schools were collapsed into one category, ISCED 3B. A slight improvement is visible from 2000 to 2006 and for the 10th and 11th grades. However, it is also evident that mean scores increased because of the improvement at the top of the achievement distribution. Looking at the vocational ISCED 3C schools it is clear that while some students caught up with their colleagues in other tracks, students performing at the lowest proficiency levels are still numerous.

Table 6: Mean achievement by PISA wave, grade and type of school program					
PISA wave:	2000	2003	2006		
Type of school program:	9th grade	9th grade	international 9th grade	national 10th grade	national 11th grade
Mean achievement	479.1	501.9	513.5	520.1	528.3
ISCED 2A *lower secondary*	-	501.9	513.5	-	-
ISCED 3A general secondary	543.4	-	-	580.8	592.6
ISCED 3A/B general, profiled	-	-	-	494.9	494.6
ISCED 3B vocational secondary	478.4	-	-	505.9	508.8
ISCED 3C vocational (basic)	357.6	-	-	388.8	384.1

An interesting comparison is between PISA 2000 results and the PISA 2006 additional "national option" sample. Table 7 gives estimates of the relative difference between achievement of students in vocational and other tracks in 2000 and in 2006 separately for the 10th and 11th grades. The results are striking. While the overall mean performance of Polish students improved significantly, the difference between students in vocational and other tracks remained almost the same, and even increased for 17-year-olds. Thus, the stratification of Polish students in the old secondary school system still exists under the new name of upper secondary schools. It seems that the reform helped to update the skills of the average student, but the negative effect of the tracking system was simply postponed by one year. The achievement gap noted in PISA 2000 is still visible and almost of the same magnitude. From one point of view this is not surprising, since the reform focused on primary and lower secondary education. On the other hand, it is now evident that the overall effect of the reform is not so positive. Intuitive claims that upper secondary education did not improve that much seems to be supported by the results presented here. Clearly, while there are visible positive effects of the reform, there are also doubts whether these positive effects are long lasting or affect all students similarly. Still, students in vocational tracks lack knowledge and skills needed to fully benefit from the modern society and economy and the reform did not change that.

Figure 5: PISA scores compared over time and with 16 and 17 year-olds

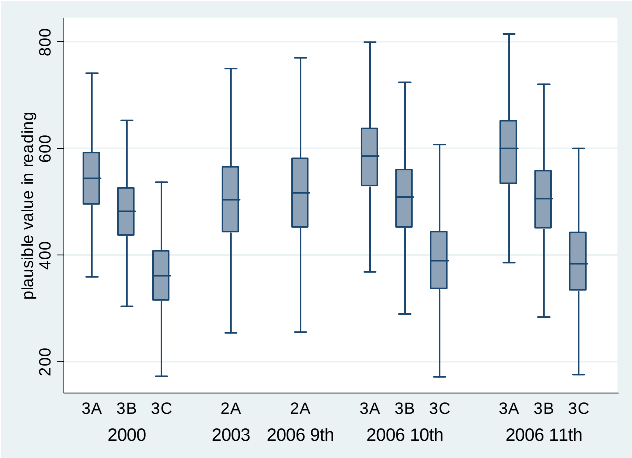


Table 7: Estimates of Relative Differences in Achievement in Vocational and Other Tracks in 2000 and 2006, and for the 10th and 11th grade special sample

	2000 9th grade	2006 10th grade	2006 11th grade
ISCED 3A + 3B	513.6	544.4	552.7
ISCED 3C	357.5	388.8	384.1
Difference	156.0	155.6	168.6
(standard error)	(7.5)	(10.2)	(10.3)

Decomposition results

Results of a simple decomposition show that, overall, two-thirds of the observed test differential between 2000 and 2006 is explained by variables in the model (Table 8). Thus, only one-third is unexplained, or in this case, due to the returns to characteristics. In terms of unexplained, most of the difference is due to returns to student characteristics, and of that, all is due to age. That is, the returns to being older increased immensely over time. In terms of the unexplained, most is due to school characteristics. Moreover, of that, almost all is due to hours of instruction. That is, receiving more than four hours per week of reading classes is associated with a higher score. The returns to hours of reading class increased over time, but much more of an increase was seen in the proportion of students that received more than four hours of language class, from 1 percent in 2000 to 76 percent in 2006.

We also present a modified decomposition (Table 9). The results are very similar. That is, most of the differential is explained, and most is due to school characteristics. Most of the differences in school characteristics are the increased hours of math and language class instructions that students were subject to because of the reform.

Table 8a: PISA Reading Scores Decomposition for Poland, PISA 2000-2006

	b2000	b2006	X2000	X2006	Determinants of Test scores Differentials			
					Endowments		*as % of total test score diff*	
Test Scores					$b_{2006}(X_{2006}-X_{2000})$	*Unexplained* $X_{2006}(b_{2006}-b_{2000})$	*Endowments*	*Unexplained*
Constant	296.5	161.49	1.00	1.00	0.00	-134.9	0.0	-205.2
Schools								
Student - teacher ratio	2.1	-0.1	12.0	11.3	0.1	-26.6	0.1	-40.5
% of certified teachers	-23.9	18.9	0.9	0.9	1.2	38.6	1.8	58.6
achievement data used	51.9	7.0	0.9	0.9	-0.4	-44.0	-0.7	-66.9
>4 hours/week reading class	3.3	42.8	0.01	0.8	32.1	0.4	48.8	0.6
attend to public school	13.9	-22.2	0.9	0.9	-0.1	-35.2	-0.2	-53.6
Student characteristics								
Age	0.3	12.9	15.7	15.7	-0.2	197.8	-0.4	300.6
Female	36.1	32.5	0.5	0.5	-0.1	-1.8	-0.1	-2.8
Family background								
Mother - Upper secondary	4.7	27.1	0.7	0.8	0.7	16.6	1.1	25.3
Mother -University	41.5	63.1	0.2	0.2	-1.5	3.6	-2.3	5.6
11–100 books	31.4	30.6	0.4	0.5	4.7	-0.3	7.2	-0.5
101-500 books	52.9	67.4	0.5	0.4	-8.0	6.8	-12.2	10.4
Computer at home	22.7	33.9	0.5	0.8	11.2	5.2	17.1	7.9
Total					39.7	26.1	60.3	39.7
Overall					65.8		100.0	

Source: Program for International Student Assessment (PISA) 2000 and 2006

Table 9: Modified Decomposition Results

	Explained (%)	Unexplained (%)
PISA Math 2000-2006		
Overall	68.7	31.3
Schools	71.0	
Family	28.9	
Student	0.2	
PISA Reading 2000-2006		
Overall	66.1	33.9
Schools	83.6	
Family	16.6	
Student	0.2	

7. Conclusions

The vocationalization of the secondary school curricula has been advocated for many decades. The call for technical and vocational schooling used to be a standard recommendation promoted by international organizations and followed by several countries. Unfortunately, the fervor for this approach has run ahead of substantial evidence on the impact of this policy. This paper contributes towards modestly filling that void by studying the effects of a change in curricular structure on educational quality.

The Polish education reform program generated an exogenous variation in vocational school attendance at the secondary school level across time that provided us the opportunity to assess the impact on test scores. Our identification strategy used the fact that likely vocational graduates did not have that option in PISA 2003, thus providing a comparison group for our empirical approach, propensity score matching and differences-in-differences estimation.

Our results suggest that vocationalization, on average, reduces test scores by a full standard deviation. Clearly, given the importance of the reform program, other factors played a role in the increase in Poland's scores in PISA. Nevertheless, the delayed vocationalization played a major role. The pathway, we argue, is through increased hours of math instruction, possibly more exposure to testing, and increased motivation on the part of students and teachers.

We substantiated our findings by taking advantage of the application of PISA to 16 and 17 year olds in PISA. We find that once vocational school options are available again, by the time students are 16, then test scores decline for those students who enter the vocational track. While this goes a long way towards proving our initial findings, it also serves as a caution to policymakers about the effectiveness of vocational schooling – when that schooling is not designed to improve math and reading skills, which we show such students can learn when they are given the opportunity, and which have become the real vocational skills in the world of work today. To increase test scores, it helps to have students study the subjects of the tests. We conclude that much of the test score increase in Poland in recent years has to do with the delayed vocationalization of the secondary school curriculum.

References

Abadie, A. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72(1): 1-19.

Attanasio, O., A. Kugler and C. Meghir. 2009. "Subsidizing Vocational Training for Disadvantaged Youth in Developing Countries: Evidence from a Randomized Trial." IZA Discussion Papers 4251, Institute for the Study of Labor (IZA).

Becker, S. and A. Ichino. 2002. "Estimation of average treatment effects based on propensity scores." *Stata Journal* (StataCorp LP) 2(4): 358-377.

Betcherman, G., K. Olivas and A. Dar. 2004. "Impacts of active labor market programs: new evidence from evaluations with particular attention to developing and transition countries." World Bank Social Protection Discussion Paper 0402.

Blinder, A. 1973. "Wage discrimination: Reduced form and structural estimates." *Journal of Human Resources 8*(4): 436–455.

Card, D.E., P. Ibarraran, F. Regalia, D. Rosas and Y. Soares. 2007. "The Labor Market Impacts of Youth Training in the Dominican Republic: Evidence from a Randomized Evaluation." NBER Working Paper No. W12883.

DiNardo, J N., M. Fortin and T. Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992." *Econometrica* 64(5): 1001- 1044.

Ekström, E. 2002. "The value of a third year in upper secondary vocational education: Evidence from a piloting scheme." Institute for Labour Market Policy Evaluation Working Paper No. 2002:23. Uppsala, Sweden.

Federowicz, M. 2007. *Badanie PISA. Umiejętności polskich gimnazjalistów*. Warszawa.

Foster, P.J. 1965. "The vocational education fallacy in development planning," in C. Anderson and M.J. Bowman, eds., *Education and Economic Development*, Chicago: Aldine.

Heckman, J.J., H. Ichimura and P. Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2): 261-94.

Heckman, J., R. Lalonde and J. Smith. 1999. "The economics and econometrics of active labor market programs," in O. Ashenfelter, D. Card, eds., *Handbook of Labor Economics*: 1865-2097.

Gruber, J. 1994. "The Incidence of Mandated Maternity Benefits." *American Economic Review* 84(3): 622-41.

Karlan, D. and M. Valdivia. 2006. "Teaching entrepreneurship: Impact of business training on microfinance clients and institutions." Yale University.

Lee, M.-J. 2005. *Micro-econometrics For Treatment-effect Analysis*. Oxford: Oxford University.

Meyer, B.D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics* 13: 151-162.

Ministry of National Education. 1998. *Reform of the education system: proposal*. Warsaw.

Murnane, R.J., J.B. Willett and F. Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77(2): 251-66.

Oaxaca, R. 1973. "Male-female wages differentials in urban labor markets." *International Economic Review 14*(3): 693–709.

OECD. 2007. *PISA 2006 Science Competencies for Tomorrow's World*. Paris.

OECD. 2003. *Literacy Skills for the World of Tomorrow - Further Results from PISA 2000.*

OECD. 2002. *PISA 2000 Technical Report*. Paris.

Psacharopoulos, G. 1997. "Vocational Education and Training Today: challenges and responses." *Journal of Vocational Education and Training* 49(3): 385-393.

Psacharopoulos, G. 1987. "To vocationalize or not to vocationalize: that is the curriculum question." *International Review of Education* 33: 187-211.

Psacharopoulos, G. and W. Loxley. 1985. *Diversified Secondary Education and Development: Evidence from Colombia and Tanzania*. Baltimore: Johns Hopkins University Press.

Rosenbaum, P.R. and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55.

Royston P. 2004. Multiple imputation of missing values. Stata Journal 4(3):227-241.

Tarozzi, A. 2007. "Calculating Comparable Statistics from Incomparable Surveys, with an Application to Poverty in India." *Journal of Business and Economic Statistics* 25(3): 314-336.

Zanutto, E. 2006. "A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data." *Journal of Data Science* 4: 67-91.