OECD Education Working Papers No. 49

The Impact of the 1999 Education Reform in Poland

OECD

https://dx.doi.org/10.1787/5kmbjgkm1m9x-en

OECD

Unclassified **EDU/WKP(2010)12**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **09-Jun-2011**

English - Or. English

DIRECTORATE FOR EDUCATION

EDU/WKP(2010)12
Unclassified

Cancels & replaces the same document of 15 July 2010

THE IMPACT OF THE 1999 EDUCATION REFORM IN POLAND

OECD Education Working Paper No. 49

This document has been prepared by Maciej Jakubowski of the OECD, Harry Anthony Patrinos and Emilo Ernesto Porta of the World Bank, and Jerzy Wisniewski of the Center for Social and Economic Research (CASE), Poland.

Contact: Andreas SCHLEICHER; E-mail: andreas.SCHLEICHER@oecd.org; Tel +33.145.24.93.66

JT03303497

Document complet disponible sur OLIS dans son format d'origine
Complete document available on OLIS in its original format

English - Or. English

OECD DIRECTORATE FOR EDUCATION

OECD EDUCATION WORKING PAPERS SERIES

This series is designed to make available to a wider readership selected studies drawing on the work of the OECD Directorate for Education. Authorship is usually collective, but principal writers are named. The papers are generally available only in their original language (English or French) with a short summary available in the other.

Comment on the series is welcome, and should be sent to either edu.contact@oecd.org or the Directorate for Education, 2, rue André Pascal, 75775 Paris CEDEX 16, France.

----

www.oecd.org/edu/workingpapers

ABSTRACT

Increasing the share of vocational secondary schooling has been a mainstay of development policy for decades, especially in formerly socialist countries. However, the transition to market economies led to significant restructuring of school systems and a decline in the number of vocational students. Exposing more students to a general curriculum could improve academic abilities. To test the hypothesis that delayed vocational streaming improves academic outcomes, this paper analyses Poland's significant improvement in international achievement tests and the restructuring of the education system, which expanded general schooling. Using propensity-score matching and difference-in-differences estimates, the authors show that delaying vocational education had a positive and significant impact on student performance on the order of one standard deviation.

RÉSUMÉ

L'expansion de l'enseignement secondaire professionnel a été un pilier de la politique de développement pendant plusieurs décennies, peut-être davantage dans les anciens pays socialistes que partout ailleurs. La transition a cependant conduit à une importante restructuration des systèmes scolaires, et notamment à une diminution de la proportion d'élèves en enseignement professionnel. L'augmentation de la proportion d'élèves inscrits en filières générales pourrait améliorer les aptitudes aux études supérieures. Cet article analyse la forte amélioration des scores obtenus par la Pologne aux tests internationaux et la restructuration du système éducatif qui a développé l'enseignement général afin de tester l'hypothèse de l'amélioration des résultats induite par une orientation plus tardive en classes de niveau. À partir d'estimations obtenues par appariement sur scores de propension et par différence de différences, les auteurs montrent que l'orientation plus tardive en filières professionnelles a eu un impact positif important, de l'ordre d'un écart-type, sur les résultats des élèves.

TABLE OF CONTENTS

THE IMPACT OF THE 1999 EDUCATION REFORM IN POLAND[1]

Introduction

1. Education policy has emphasised vocational training since the Second World War. It is often argued that vocational skills are necessary to create jobs, employment and productivity. Logically, a country needs vocational education to equip its workers with the technical skills needed for the country to modernise and develop economically. Psacharopoulos (1997) summarises the reasons for increasing the proportion of students in vocational education programmes as follows:

i. **Youth unemployment**: With one step, policy makers can take youth off the streets and at the same time equip them with skills that could be used later in the labour market.

ii. **Instilling technological knowledge**: Since the Industrial Revolution, it has been commonly believed that economic progress depends on technological knowhow. Given that assumption, vocational education must expand.

iii. **Academically less able students**: Students who are "unable" to advance through the school system, especially the academic curriculum of secondary education, have been a constant concern. In theory, giving them access to vocational education would equip them with the skills to do something useful later in life.

iv. **Lack of mid-level technicians**. All countries suffer from a "scarce" supply of skilled workers, such as plumbers and nurses. It would therefore seem logical to create vocational schools and training institutions to provide a labour force with these specialised skills.

v. **Poverty among urban dwellers**: Given the increased poverty of urban dwellers, providing vocational education would give useful skills to unemployed people and help them find jobs to raise their incomes.

vi. **Economic globalisation**: The advent of free trade and the rise of multinationals have implications for the kinds of vocational education provided to the labour force.

2. Since the Second World War, many countries have developed vocational education systems. Socialist countries integrated vocational schooling into the overall economic planning system, assigning them to different ministries. In these models, employment was guaranteed. However, once the transition to a market economy began, the link between vocational education and employment was broken, leaving vocational students without jobs and without the skills demanded by the labour market.

1 . 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, OECD and the World Bank. The views expressed here are those of the authors and should not be attributed to the OECD and the World Bank Group. Address all correspondence to Maciej Jakubowski at maciej.jakubowski@oecd.org

3. Indeed, the emphasis on vocational education has been under attack for many decades. Psacharopoulos (1987) argues that the social costs of vocational education may not match the social benefits associated with it. The argument that vocational education would bring industrialisation and jobs was challenged early on by Foster (1965), who called it the "vocational school fallacy". More important, the vocational skills of today–what is needed in the world of work, what students must learn to compete–are not the traditional skills linked to specific jobs; rather, they are the skills of critical thinking and "learning to learn" (see Murnane *et al*. 1995) that are exemplified by success in mathematics, reading and science, for example.

4. Despite its prominent place in school policy, there has been little rigorous evaluation of the education vocational schools provide. Much more work has focused on financing, arguing that general skills are a public priority while specific vocational skills should be privately financed or financed by employers (Becker 1964). Wage effects or returns to schooling for vocational tracks have been estimated and compared to general or academic tracks. Overall, cost-benefit studies show that returns are lower and costs higher (Psacharopoulos and Patrinos 2004).

5. Some empirical literature suggests that that there are advantages to targeted vocational *training* programmes that are not school-based (Karlan and Valdivia 2006). Evaluations of the randomised training programmes in the United States show modest effects, at best (see, for example, Heckman, Lalonde and Smith 1999). Evidence of the effectiveness of training in developing countries is more limited. Betcherman, Olivas and Dar (2004) review 69 impact evaluations of unemployed and youth training programmes, only 19 of which are in developing countries. They find that the impacts in developing countries are more positive than the impacts of programmes in the United States and Europe. Most of those programmes, however, are not experimental. Card et al. (2007) report on the first randomised evaluation of a job-training programme in Latin America. The subsidised programme in the Dominican Republic showed no impact on employment, a marginally significant impact on hourly wages and on the probability of health insurance coverage, conditional on employment. Attanasio, Kugler and Meghir (2009) evaluate the impact on employment and earnings outcomes of a randomised training programme for disadvantaged youth in Colombia. They find that the programme raises earnings and employment for both men and women, with greater impact on women. Cost-benefit analysis of these results suggests that the programme generates a large net gain, especially for women.

6. Fewer evaluations, randomised or otherwise, have been undertaken on the impacts of vocational education. Earlier assessments of vocational education programmes in a number of countries, including Colombia and Tanzania, have shown that most graduates of such schools go to university rather than entering manual occupations (Psacharopoulos and Loxley 1985). In 1991, Sweden's upper secondary school two-year vocational programmes were transformed into three-year programmes 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 secondary education, university enrolment, and the rate of inactivity. Results suggest positive effects on upper secondary education for those who lived in a pilot municipality in 1990. One of the important changes was that the third year of upper secondary vocational education gave individuals the skills needed to continue to higher education. However, the third year did not have a statistically significant effect on the probability of continuing to higher education, at least not within six years after completing upper secondary education (Ekström 2002). To our knowledge, no rigorous study has been undertaken on the learning outcomes associated with vocational secondary schooling.

7. Poland is a good case for such an evaluation. 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 old eight-year primary school that was followed by early 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. This structural change was accompanied by curricular reform. A concept of core curricula was developed that aimed to provide schools with extensive autonomy and responsibility. A system of examinations and tests at the end of primary and lower secondary was also introduced.

8. The purpose of our paper is to explain Poland's significant improvement in international achievement tests in recent years. 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 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.

9. 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 difference-in-differences to create counterfactual scores for the group of likely vocational students in subsequent years–and the OECD's Programme for International Student Assessment (PISA), an internationally comparable standardised student test conducted every three years to test reading, mathematics and science achievement of 15-year-olds. We use PISA data from 2000, 2003 and 2006, with 2000 as the baseline, since most of the existing students were continuing their lower secondary schooling under the old 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 early vocational education.

10. This paper is composed of eight sections: 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; and we summarise our conclusions and discuss the policy implications in section 8.

Reform of 1998-1999

11. In 1998, the Polish Minister of Education presented the outline of the reform, setting the following goals (Ministry of National Education 1998):

1. Raise the level of education in society by increasing the number of people with secondary and higher education qualifications;
2. Ensure equal educational opportunities; and
3. Support improvements in the quality of education.

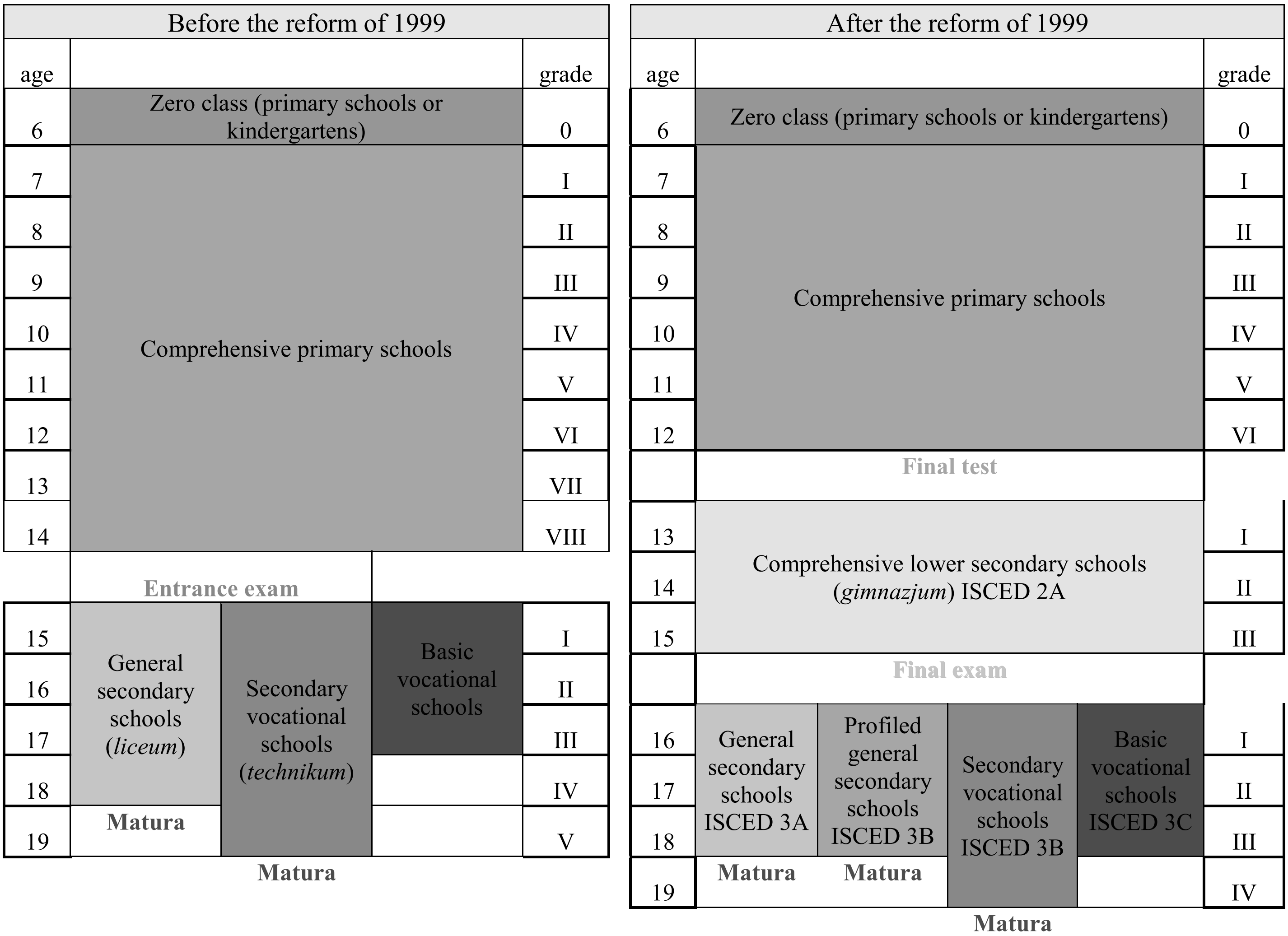
12. The reform was envisaged to cover:

- the structure of the education system, ranging from nursery school to doctoral studies; this included re-structuring the entire system;
- administration and supervision methods;

- the curriculum, including introducing a core curriculum and changing the way teaching is organised and provided;
- an independent assessment and examination system;
- school finance; and
- teacher qualifications, which would be linked with their promotion paths, and the remuneration system.

13. The structural changes resulted in a new type of school: the lower secondary school "gymnasium", which became a symbol of the reform. The previous structure, comprising the eight-year primary school followed by the four-year secondary school or the three-year vocational school, would be replaced by a system described as 6+3+3 (Figure 1). This meant that education in the primary school would be reduced to six years. A pupil would then continue his/her education in a three-year gymnasium. Only after completing three years in the gymnasium would he/she move on to a three-year secondary school (specialised lyceum) or a two-year vocational school. The reform postponed for one year the choice between the secondary-level general or vocational curriculum. With these stages in education now clearly defined, pupil achievements could be reliably assessed through tests and examinations.

Figure 1: Structure of the Polish Education System



14. The reformers assumed that the gymnasia would allow Poland to raise the level of education, particularly in rural areas where the schools were small. The new lower secondary schools would be larger, with at least 150 pupils. They would also be well-equipped and would employ teachers with adequate qualifications. Since the number of pupils in the school varies with the school-catchment area, establishing the gymnasia involved reorganising the school network. The structural reform did not cover nursery schools and did not result in lowering the age at which compulsory schooling begins (7 years).

15. Reformers had two main arguments for the changes. First, dividing education into stages would allow teaching methods and curricula to better meet the specific needs of pupils of various ages. Second, a structural reform would have to be linked with a curricular reform, otherwise those teachers who resisted the reform may continue to teach their pupils in the same ways as they had for many years. So teachers were encouraged to change what they taught and how they taught it.

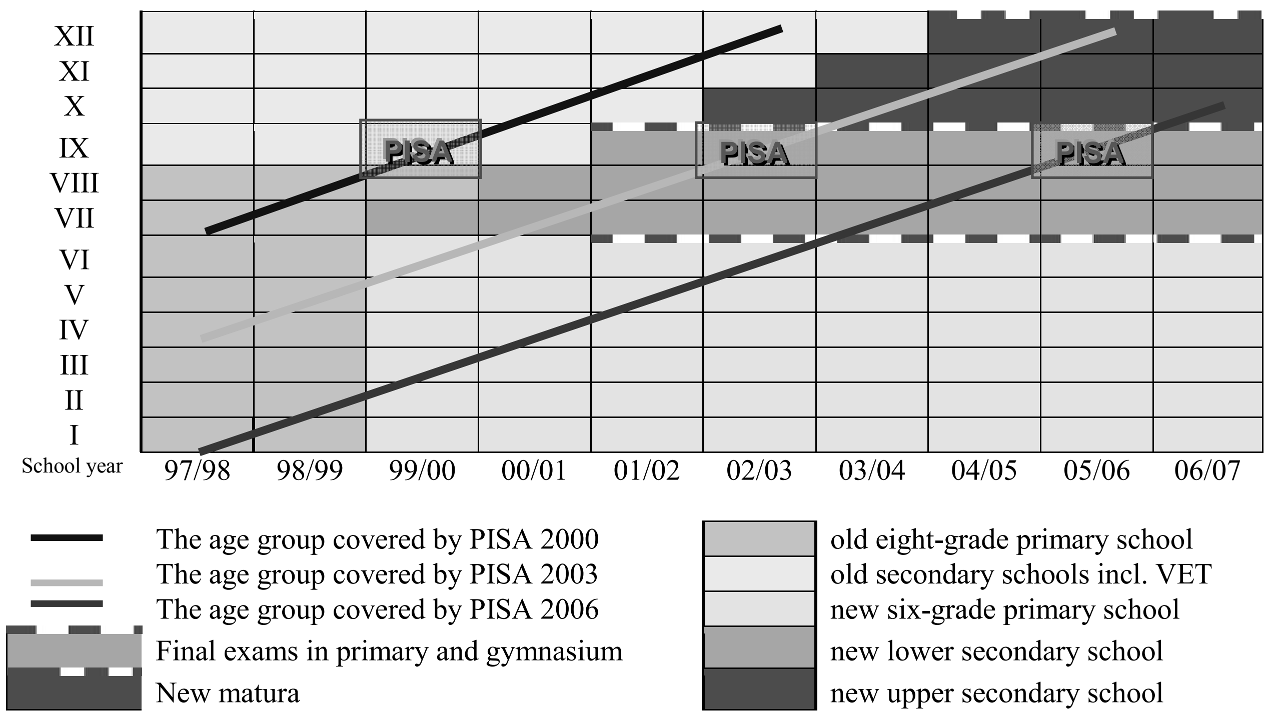
16. After years of complaints about overloaded curricula and disputes about the way forward, the concept of core curricula was adopted. The concept aimed to offer schools extensive autonomy and

responsibility. Schools were to build their own curricula within a pre-determined general framework while balancing the three goals of education: imparting knowledge, developing skills and shaping attitudes. The curricular reform was designed not only to change the content of school education and encourage innovative teaching methods, but also to change the teaching philosophy and culture of schools. Instead of passively following the instructions of the educational authorities, teachers were expected to develop their own teaching styles, which would be tailored to the needs of their pupils.

17. Introducing curricular reform based on decentralisation required implementation of a system for collecting information and monitoring the education system at the same time. Reformers thus decided to organise compulsory tests to assess pupil achievements at the end of the primary and lower-secondary cycles. Both of these were administered for the first time in 2002. Schooling would culminate with the *matura* examination, taken at the end of upper-secondary education.All these examinations were to be organised, set and corrected by the central examination board and regional examination boards, new institutions set up as part of the reform. The *matura* was administered for the first time in 2005. The results of the primary school test do not affect the students' school career, as the completion of the cycle does not depend on the results. In the selection process for upper-secondary schools, the score earned on the gymnasium final exam is considered together with the pupil's final marks.

18. The age cohorts covered by PISA in 2000, 2003 and 2006 have been affected by the reform in different ways (Figure 2). 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 started their education in primary school in the former system but attended the gymnasia, the flagship of the reform. They did not take the final test in the sixth grade of primary school. The test was administered for the first time in 2002, when they were already gymnasium students. The group covered by PISA 2006 had been part of the reformed educational system for most of their school careers. 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.

Figure 2: PISA and the reform cohorts



19. The group covered by PISA 2000 consisted of the first grade students of the pre-reform secondary schools: general lyceum, which students could enter only if they passed an entrance exam, secondary vocational school and basic vocational school, which was 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 compulsory gymnasium, so the results showed smaller variations among schools and larger ones among students within schools.

20. Among the PISA 2000 participants, only students of lyceums and some secondary vocational schools had previous experience in taking a written entrance exam. The others had no experience at all. The lyceum entrance exam was not, in fact, a test: it consisted of a written essay and five slightly complicated, but standard, mathematical problems. The first national final tests after primary school and gymnasium were carried out in 2002. At that time, the group of PISA 2003 were in the second grade of gymnasium, so they did not take the final primary school test; however, the PISA 2006 group were then still in the fifth grade of primary school, so they took the full set of the new external exams.

21. 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 written 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 for their upcoming final gymnasium exams.

22. PISA 2006 participants were well acquainted with doing tests. They took the final primary school test and had three years of preparation for the gymnasium exam. Konarzewski (2004) shows that teachers took the 2002 final exams, the first of their kind, very seriously. One-third of teachers in a representative sampling said that they changed their teaching to familiarise students with test requirements. Testing was also considered when choosing textbooks and other supporting teaching materials. Twenty-six percent of the teachers said that unsatisfactory test results were not caused by students' poor knowledge or low skills, but by their lack of experience in taking such tests. Teachers thus concluded that it was important to practice taking tests. Konarzewski (2008) shows that a substantial amount of time is devoted to solving test-type problems and doing mock exams in all gymnasia. Some five percent of the respondents have changed their assessment schemes, making them more test-like. In his conclusion, Konarzewski (2008) writes: "The test exam, being so predictable as ours, each year less and less measures the competences of gymnasium leavers but more and more the effort and time spent by schools on training students to do the exams."

Relative increase in scores

23. 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 (see Table 1). Reading scores have steadily improved over time, from 479, to 497, to 508 in the latest round. 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. In science, the scores are 483, 498 and 498.

Table 1: Top 10 reading over time, PISA

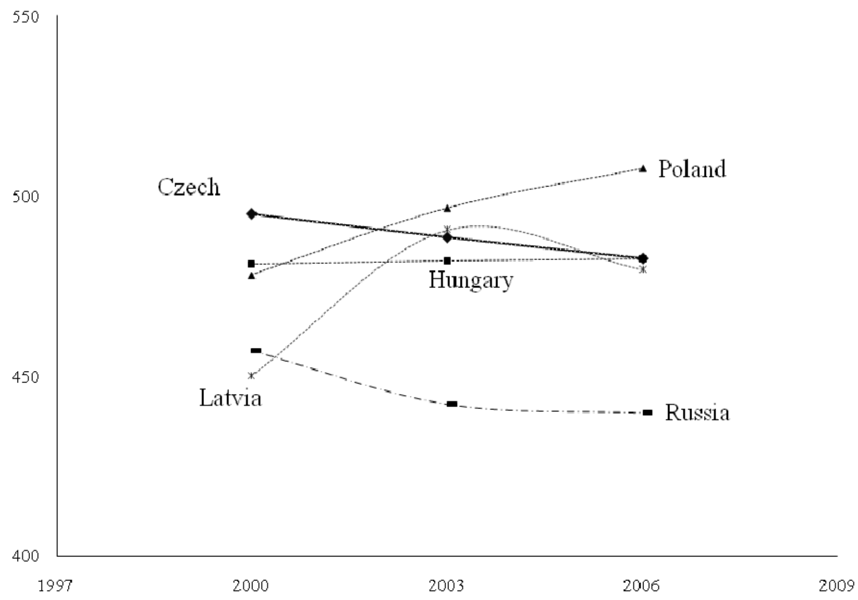
	2000		2003		2006	
1	Finland	549	Finland	543	Korea	556
2	Netherlands	537	Korea	534	Finland	547
3	Canada	535	Canada	528	Hong Kong	536
4	Hong Kong	532	Australia	525	Canada	527
5	Australia	528	Liechtenstein	525	New Zealand	521
6	Ireland	528	New Zealand	522	Ireland	517
7	New Zealand	526	Ireland	515	Australia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

Hypotheses for explaining change over time

24. While several factors could explain these changes, it is difficult to find causal relationships. To assess the effectiveness of national education policies, only samples that contain similar student and parent profiles can be compared internationally. For example, if two countries differ in levels of parental education, which strongly affects student outcomes, then it is not valid to compare mean performance in these two countries as a way of determining whether one has a more effective education policy than the other. It is most likely that the difference in mean performance depends more on the difference in parental education than on the policy itself. Thus, any comparison of unadjusted samples could be irrelevant or unhelpful to policy makers. Similarly, to compare achievement levels in a particular country in different years, the samples have to be adjusted to make them fully comparable. While PISA organisers 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.

25. *Not all transition countries improved over time*. Figure 3 shows the performance of the five Eastern European countries that participated in all three rounds of PISA. Poland is the only country with consistent improvement over time. In fact, among the five countries that participated in all three rounds of PISA, only Latvia and Poland improved over time. 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.

Figure 3: PISA Reading Performance in ECA over time



26. *Reform led to improvement.* We compare changes in student performance in Poland across 2000, 2003 and 2006. We show that improvement in student scores is due to the delay of streaming into vocational tracks and to greater resources devoted to education, particularly to instruction time.

27. *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.

Empirical methods and data

28. We test whether the reform—specifically, the change in the structure of the school system–led to the improvement in test scores by delaying vocational education. 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.

29. 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 known as the "curse of dimensionality". To resolve this problem, propensity-score matching methods were proposed by Rosenbaum and Rubin. In these methods, instead of matching multiple characteristics the *propensity score* is balanced across comparison groups.

30. 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:

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

31. Tarozzi (2007) argues that such reweighting produces comparable outcome distributions. *Depvar*=1 is defined as being in 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. PISA survey design was accounted for by multiplying propensity-score weights and survey weights.

32. As covariates, we used gender, age, mother's 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, the number of migrants in the Polish sample is negligible. Missing data were imputed using the multiple imputation approach (Royston 2004). Results without any imputation were qualitatively similar, though less precise because of smaller sample sizes.

Estimates of score change for students in different tracks

33. 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 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 change in test scores cannot be solely attributed to replacing the traditional secondary school tracks with lower secondary schools for 15-year-olds.

34. 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. The point is: without the reform, 15-year-old students in vocational schools would not have had the opportunity to study in general programmes; 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.

35. Our approach to estimating the differential score change is similar to the difference-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. 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.

36. We use treatment-evaluation nomenclature (see Lee 2005) to formally define the groups. The treatment is defined as 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 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 econometric evaluation 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.

37. 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 difference-in-differences model is formulated as:

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

38. 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). 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 difference-in-differences estimator is usually employed that controls for the set of covariates:

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

39. The crucial assumption here is that quasi-experimental groups differ only by observable covariates. This condition eliminates any bias. Typically, the difference-in-differences 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 (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).

40. 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.

41. As already mentioned, 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 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 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} = $P(T\text{=}2000|X)$, 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_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.

42. 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:

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

43. In this equation, $E(Y^1 \mid D=1)$ and $E(Y^1 \mid D=0)$ are directly observed in the data, but $E(Y^0 \mid P_T^{2000}, D=1)$ and $E(Y^0 \mid P_C^{2000}, D=0)$ 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 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 difference-in-differences 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.

44. Propensity scores were estimated using logit regressions. Two kinds of propensity score matching were then employed: 1-to-1 nearest neighbour 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.

45. 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 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 the PISA survey, and family structure, are 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 ensure that the sample size and performance distribution are untouched by the matching exercise, missing values for matching covariates were imputed through multiple imputation models (Royston 2004).

46. The PISA survey has a complex structure, similar to methods commonly used in other educational surveys, such as the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), or the United States' National Assessment of Educational Progress (NAEP), 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 (see Zanutto 2006, for example, of analysis with survey weights and stratification matching). We used 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 are answering randomly assigned groups of test items, so-called booklets, but responses are put into one common scale using psychometric 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 of repeating each analysis five times, with each plausible value used once to allow for measurement error in student performance. 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 (see OECD 2002, 2005).

47. The final set of variables from the PISA dataset 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 absence of bias in 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 groups to mimic re-sampling of the original population. Replicate weights are alternative sample weights that represent a sub-sample based on the original sampling design. PISA provides replicate weights compatible with Fay's adjusted Balanced Repeated Replication. These weights were constructed to reflect the sampling design, including any country-specific modifications, as well as non-response by students or schools (OECD 2002: 89-98). Standard errors were obtained by the BRR method. For us, the additional benefit of using BRR weights is that these were produced by survey organisers who used confidential information not available to external users.

Decompose change over time

48. In order to try to explain how the reform may have resulted in improved student achievement, we perform a simple decomposition analysis. We decompose reading scores between 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 (Hanushek 1986, 2002; Todd and Wolpin 2003; Glewwe 2002). Education production function is a model that relates various inputs affecting student learning, such as learning time or family resources, to measured outputs. In this case, the measured outputs are the PISA standardised reading test scores.

49. Past research is inconclusive about which school and family characteristics, such as class size, teacher experience, teacher education and mother's employment, influence students' achievement. Although achievement in education largely depends on the individual child's efforts and inherent capacities, a large body of evidence supports the theory that family background influences student outcomes (Fertig 2003; Fertig and Schmidt 2002; Currie and Thomas 1999). Consequently, researchers must control for individual pupil characteristics as well as for family background, and for characteristics of

the school environment and the education system. Evidence also suggests that socio-economic and family background variables, such as parents' education and the number of books in the household, are important determinants of test scores at early ages (Fryer and Levitt 2002). We thus specify and estimate education production functions that relate students' achievement to individual, family and school inputs. We then decompose the over-time test-score gap into an explained component, accounting for student, family, and school characteristics, and an "unexplained" component–or returns, the efficiency with which the country can convert characteristics into student learning outcomes as measured by test scores–using the traditional Oaxaca (1973)-Blinder (1973) decomposition method. The education production functions were estimated by linear regressions accounting for clustering of students at the school level.

50. The model specification for estimating the production function for cognitive achievement is:

$$T_{ija} = T_a(A_{ija}, F_{ija}, S_{ija}) + \epsilon_{ija}$$

where T_{ij} is the observed test score (from PISA reading) of student i in household j at time a (time of the test), A_{ija} is a vector of individual student characteristics, F_{ija} is a vector of parent inputs, S_{ija} is a vector of school-related inputs, and ϵ_{ija} is an additive error, which includes all the omitted variables, including those that relate to the history of past inputs, endowed mental capacity and measurement error. The linear specification, after dropping subscript a, of the production function is given by:

$$T_{ij} = \beta_0 + \beta_1 A_{ij} + \beta_2 F_{ij} + \beta_3 S_{ij} + \epsilon_{ij}$$

where β0 to β4 are coefficients to be estimated. The standard procedure for analysing the determinants of the test score differences over time is to fit equations between test scores and observed characteristics. The observed test score differential can be decomposed as:

$$T_{2006} – T_{2000} = (X_{2006} - X_{2000})\beta_{2006} + X_{2006}(\beta_{2006} - \beta_{2000})$$

where T is the standardised test score, Xi is a vector of student, family and school characteristics for the ith individual, β is a vector of coefficients, and 2006, 2000 subscripts are identifiers of the PISA test score in reading in years 2000 and 2006, evaluated at 2006 values.

51. The overall test-score increase can thus be decomposed into two components: one is the portion attributed to differences in characteristics ($X_{2006} – X_{2000}$) evaluated with the 2006 values, or 2006 group performance (β_{2006}); the other portion is attributable to differences in effects on performance (β_{2006} - β_{2000}) of 2000 and 2006 students derived from the same characteristics. This second, unexplained component, while more difficult to interpret in this context compared to an earnings gap decomposition framework, can be assigned more than one interpretation. For example, the unexplained portion of the test-score increase may reflect certain unobserved family characteristics that are correlated with achievement over time, possibly relating to household wealth. In addition, it may be that the different cohorts of students do not reap the same benefits from equivalent school and classroom resources.The unexplained component may also reflect the impact of changes over time based on past reforms that both increased school enrolments in Poland and helped improve the quality of school inputs. Some of the above coefficient estimates may be subject to biases. For example, if a school characteristic is correlated with unobserved family characteristics that influence achievement, such as family wealth and parents' motivation, the effect of attending a school with such characteristics may be biased.

Results

52. Our analysis focuses on reading literacy, as 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 ninth grade in Poland. In PISA 2000, only the ninth grade was sampled; in PISA 2003 and 2006, students from the seventh, eighth and tenth grades were also sampled. The results suggest that students 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.

53. Reweighting clearly lowers the mean scores of students in 2003 and 2006 (Table 2) 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, 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 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 and 2006 results for Poland in reading factual (with 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

54. While the change in mean scores is interesting, looking at the change in whole distributions gives a more detailed picture. Figures 4 and 5 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 percent of students scored in the top two reading proficiency levels, the fourth and fifth levels, compared to the OECD average of 31.8 percent. In 2006, this percentage increased to 34.7 percent, compared to the OECD average of 29.3 percent. Meanwhile, the percentage of Polish students below or at the first proficiency level was 23.3 percent in 2000, compared to the OECD average of 17.9 percent, and 16.2 percent in 2006, compared to the OECD average of 20.1 percent (OECD 2003: Table 2.1a; OECD 2007: Table 6.1a). 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?

Figure 4: Change in reading literacy distribution between PISA 2000 and 2006

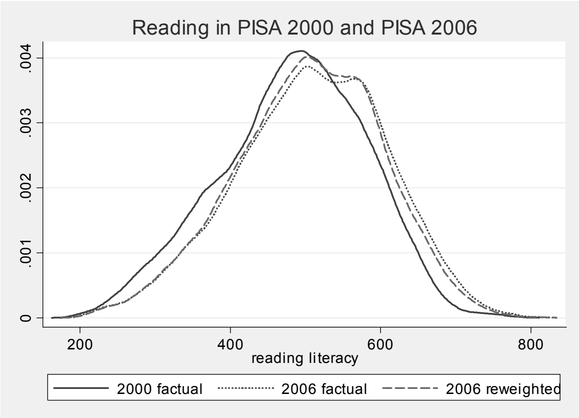
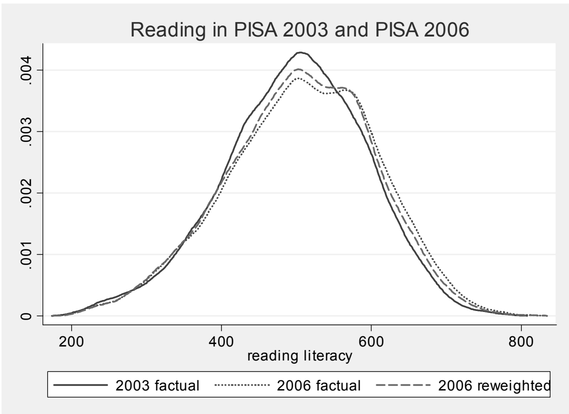


Figure 5: Change in reading literacy distribution between PISA 2003 and 2006



Estimates of score change for students in different tracks

55. Results for difference-in-differences propensity score-matching estimates of the effect of abolishing the tracking system for 15-year-olds in Poland 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. Results for students in vocational and non-vocational tracks are also 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, as described above.

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

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

57. Results are summarised in Table 4, which shows the estimates of score improvement.[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. Again, there is 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 concern the hypothetical performance improvement from 2000 in different tracks. Performance improvement for potential students of former vocational schools is 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. 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 reform. However, the benefits for students in other tracks are not that evident. Students in mixed-general schools improved their scores only slightly in 2003 but noticeably in 2006. Students in the general track would potentially have lower scores in 2003 and similar performance in 2006.

2. The numbers presented in the third row, after the name of the comparison and matching method, show 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 students 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 accounts for complex survey design (stratification, clustering, and response adjustments).

58. 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 teachers adjusted their methods to suit more diverse classrooms or 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 in PISA. 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 much more time learning non-vocational subjects. What is striking is the magnitude of the improvement–nearly one standard deviation of PISA international scores–and the speed with which students adapted to the new system. Clearly, adding just a few months of comprehensive education in the place of vocational education dramatically changes the general skills for a large number of students.

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.

59. Relevant difference-in-differences estimates of performance change for vocational school students 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 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 (100). Relative improvement in comparison to students in mixed general-vocational schools is slightly lower but still substantial.

Table 5: Relative score change (difference-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

60. 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 and mixed general-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.

Additional analyses

61. 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 (see Federowicz 2007 for the report on PISA 2006 in Poland). 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. In other words, knowing the students' achievement at the end of lower secondary schools, we can determine to what extent students updated their skills in the different types of upper secondary schools.

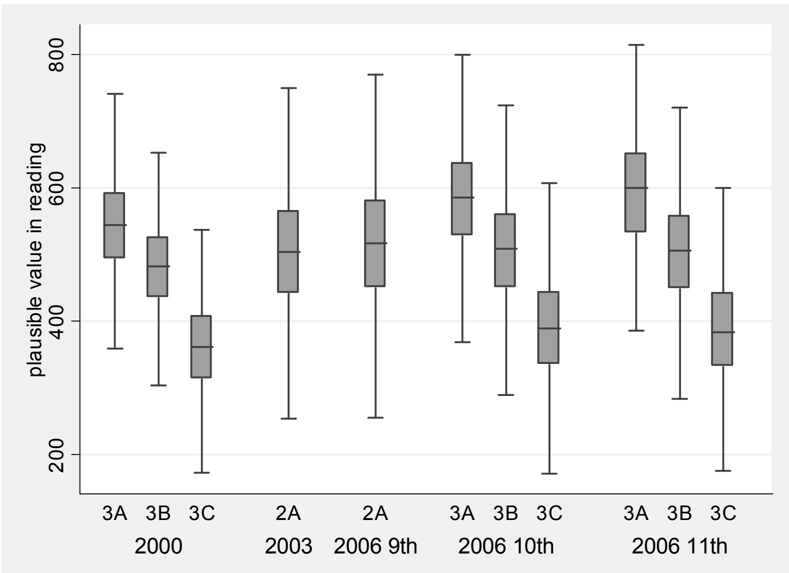
Analysis of PISA 2006 "national option" samples

62. Estimates of mean achievement by PISA cycle, grade and type of school programme are presented in Table 6. First, 16-year-old students in the tenth grade score, on average, higher than do 15-year-olds in the ninth grade, and 17-year-olds in the eleventh grade score higher than 16-year-olds. This is in line with intuition that older students are better able to pass PISA tests. However, when we look at the type of school programmes, it is clear that mainly students in ISCED 3A schools improved, while 17-year-old students in vocational schools had even lower scores. This seems to be counterintuitive, but there are two highly likely explanations. First, students change tracks, mostly in the tenth 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 ISCED 3C tracks devote more time to vocational training in higher grades, their general skills, tested in PISA, could decline. Consequently, slightly lower achievement in ISCED 3C is not that surprising.

Table 6: Mean achievement by PISA wave, grade and type of school programme					
PISA wave:	2000	2003	2006		
Type of school programme:	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 school	-	501.9	513.5	-	-
ISCED 3A general secondary	543.4	-	-	580.8	592.6
ISCED 3A/B general, profiled secondary	-	-	-	494.9	494.6
ISCED 3B vocational secondary	478.4	-	-	505.9	508.8
ISCED 3C vocational (basic)	357.6	-	-	388.8	384.1

63. Box plots presented below summarise score distribution for the categories presented in Table 6 (Figure 6). This time, data for vocational upper-secondary schools and general (mixed) upper-secondary schools were collapsed into one category, ISCED 3B. A slight improvement is seen from 2000 to 2006 and for the tenth and eleventh grades. However, it is also evident that mean scores increased because of the improvement at the top of the achievement distribution. Among the vocational ISCED 3C schools, it is clear that while some students caught up with their colleagues in other tracks, most students performed at the lowest proficiency levels.

Figure 6: PISA scores compared over time and with 16 and 17-year-olds



64. 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 tenth and eleventh 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 remains under the new name of upper secondary schools.

65. 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 evident and almost of the same magnitude. On the one hand, 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 these results. While the positive effects of the reform are evident, there are also doubts as to whether these effects are long lasting or affect all students in the same way. Still, students in vocational tracks lack the knowledge and skills needed to fully benefit from the modern society and economy, and the reform did not change that.

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

66. We present the decomposition results in order to explain one of the ways the reform may have led to improved student achievement. Tables 8a and 8b present the results of production-function estimates along with the decomposition results in reading. Overall, two-thirds of the observed test-score differential between PISA 2000 and 2006 is explained by 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 percent in 2000 to 76 percent in 2006. At the student level, the change in the effect of student age has a large impact on the overall performance change. That is, the positive effect of being older increased over time.

Table 8a: PISA reading scores decomposition for Poland, PISA 2000-2006

Test Scores	b2000	b2006	X2000	X2006	Determinants of Test scores Differentials			
							as % of total test score diff	
					Endowments	*Unexplained*	*Endowments*	*Unexplained*
					$b_{2006}(X_{2006}-X_{2000})$	$X_{2006}(b_{2006}-b_{2000})$		
Constant	296.47	161.49	1.00	1.00	0.00	-134.98	0.0	-205.2
Schools								
Student - teacher ratio	2.08	-0.14	12.01	11.33	0.09	-26.61	0.1	-40.5
% of certified teachers	-23.92	18.85	0.90	0.97	1.21	38.57	1.8	58.6
Achievement data used to evaluated teachers and principal performance	51.94	7.00	0.98	0.92	-0.43	-44.01	-0.7	-66.9
More than 4 hours per week of language class	3.27	42.77	0.01	0.76	32.09	0.41	48.8	0.6
attend to public school	13.89	-22.18	0.98	0.98	-0.13	-35.25	-0.2	-53.6
Student characteristics								
Age	0.28	12.85	15.73	15.71	-0.23	197.76	-0.4	300.6
Female	36.12	32.53	0.51	0.51	-0.05	-1.83	-0.1	-2.8
Family background								
Mother - upper secondary	4.68	27.11	0.74	0.77	0.70	16.65	1.1	25.3
Mother -university	41.49	63.09	0.17	0.15	-1.52	3.65	-2.3	5.6
11–100 books	31.38	30.58	0.39	0.54	4.75	-0.31	7.2	-0.5
101-500 books	52.90	67.39	0.47	0.35	-8.03	6.87	-12.2	10.4
Computer at home	22.74	33.89	0.47	0.80	11.22	5.19	17.1	7.9
Total					39.7	26.1	60.3	39.7
Overall					65.8		100.0	

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

Table 8b: Determinants of PISA differentials, reading 2000-2006

	as % of total test score diff	
	Endowments	*Unexplained*
Constant	0.0	-205.2
Schools	49.9	-101.7
Family	10.8	48.7
Student	-0.4	297.9
Total	60.3	39.7
Overall	100	

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

67. The results are similar in the modified decomposition (Table 9). Most of the differential can be explained by school characteristics, particularly the increase in class hours for language instruction that was part of the reform.

Table 9: Modified decomposition results

	Explained (%)	Unexplained (%)
PISA Reading 2000-2006		
Overall	66.1	33.9
Schools	83.6	
Family	16.6	
Student	0.2	

Conclusions

68. Including more vocational training in 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 organisations and implemented by several countries. Unfortunately, the enthusiasm for this approach was not based on any substantial evidence of its benefits to students.

69. The Polish education reform programme gave us the opportunity to assess the impact of vocational training on test scores. Our identification strategy was based on the fact that likely vocational graduates did not have that option in PISA 2003, which provided a comparison group for our empirical approach, propensity-score matching and difference-in-differences estimation.

70. Our results suggest that, on average, vocational schooling reduces test scores by a full standard deviation. While other aspects of the reform programme no doubt helped improve Poland's PISA scores, delayed entry into vocational education played a major role. We argue that the way to achieve better PISA scores is through more hours of instruction, greater exposure to testing, and increased student and teacher motivation.

71. 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. 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 that schooling 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.

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, G. (1964), *Human Capital*, The University of Chicago Press, Chicago.

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.

Currie, J. and D. Thomas (1999), "Early test scores, socioeconomic status, and future outcomes", *NBER Working Paper* No. 6943.

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, IFiS PAN, Warszawa 2007.

Fertig, M. (2003), "Who is to blame? The determinants of German students' achievement in the PISA 2000 study", *IZA Discussion Paper* No. 739.

Fertig, M. and C.M. Schmidt (2002), "The role of background factors for reading literacy: Straight national scores in the PISA 2000 study", *IZA Discussion Paper* No. 545.

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

Fryer, R. and S. Levitt (2002), "Understanding the black-white test-score gap in the first two years of school", *NBER Working Paper* No. 8975.

Glewwe, P. (2002), "Schools and skills in developing countries: Education policies and socioeconomic outcomes", *Journal of Economic Literature* 40(2): 436-82.

Hanushek, E. (1986), "The economics of schooling: Production and efficiency in public schools", *Journal of Economic Literature* 24(3): 1141-1177.

Hanushek, E. (2002), "Publicly provided education", in A.J. Auerbach and M. Feldstein (eds.), *Handbook of public economics* (vol. 4). Elsevier, Amsterdam.

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 and D. Card (eds.), *Handbook of Labor Economics* (vol. 3A): 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, New Haven.

Konarzewski, K. (2008), *Przygotowanie uczniów do egzaminu: pokusa łatwego zysku*, Raport badawczy, ISP, Warszawa.

Konarzewski, K. (2004), *Reforma oświaty, podstawa programowa i warunki kształcenia*, ISP, Warszawa.

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

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*, WSiP, 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*, Vols. 1 and 2, Paris.

OECD (2005), *PISA 2003: Technical Report*, Paris.

OECD (2003), *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*, Paris.

OECD (2002) *PISA 2000: Technical Report*, Paris.

OECD (2001), *Knowledge and Skills for Life: First Results from PISA 2000*, 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*, Johns Hopkins University Press, Baltimore.

Psacharopoulos, G. and H.A. Patrinos (2004), "Returns to investment in education: A further update", *Education Economics* 12(2): 111-134.

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.

Todd, P.E. and I. Wolpin (2003), "On the specification and estimation of the production function for cognitive achievement." *Economic Journal* 113: F3-F33.

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