Grade retention and educational attainment

Exploiting the 2001 Reform by the French-Speaking Community of Belgium

and Synthetic Control Methods

Michèle Belot

Oxford University, Nuffield Centre for Experimental Social Sciences (CESS), Nuffield College

michele.belot@nuffield.ox.ac.uk

Vincent Vandenberghe

Université catholique de Louvain (UCL), Economics School of Louvain (ESL)

vincent.vandenberghe@uclouvain.be

Abstract

This paper evaluates the effects of grade retention on attainment by exploiting a reform introduced in 2001 in the French-Speaking Community of Belgium whereby the possibility of grade retention in grade 7 was reintroduced. It uses the Synthetic Control Method to identify the best possible pre-treatment control. Data come from three waves of the PISA study (corresponding to periods before and after the reform) that contains test scores of representative samples of 15 year-olds. These are used essentially to answer two questions. First, has the 2001 grade repetition reform at least succeeded at filtering out weaker pupils, pupils who would presumably be disadvantaged by being promoted directly to higher grades. This is a minimum condition for grade retention to be justifiable. Second, do these "treated" students achieve better/worse when they repeat (and attend a lower grade) than when they are "socially promoted" (and attend the age 15 reference grade 10)? We find significant evidence of positive screening but we fail to demonstrate that those filtered out perform differently under the "grade repetition" regime than under the "social promotion" regime.

JEL: I20, I28, H52 **Keywords**: Grade retention, educational attainment, synthetic control method

1. Introduction

Grade retention (or repetition) is the object of an ongoing debate in many developed countries. Some countries privilege a system of "social promotion", which allows pupils to be promoted to higher grades independently of their performance, while other countries have instituted more or less strict policies of grade retention, conditioning promotion to higher grades on educational achievements. As a consequence, there is a considerable variation in grade retention rates[1] across OECD countries (Figure 1). Countries/entities like the Netherlands, Austria, Portugal and the French-Speaking Community of Belgium have relatively high rates of grade retention (going up to 50% of pupils having repeated a year or more by the time they reach the end of compulsory schooling); while countries like Denmark, Sweden, Japan, Norway and the UK have no grade retention at all.

Grade retention imposes a cost on society, both in terms of the opportunity costs of those pupils who are forced to repeat a year, but also in terms of teaching resources. Indeed, grade retention often implies larger class sizes and more pressure on the (limited) teaching resources. Overall pedagogues do not generally support the effectiveness of grade retention and the ensuing differences in the grade attended by pupils (McCoy and Reynolds, 1999). They argue that grade retention has negative effects on self-esteem and academic performance, and even on non-academic outcomes such as crime and teenage pregnancy. On the other hand, the proponents of grade retention argue that it may have motivational effects on pupils – the threat of being retained playing the role of a "stick".

There is a large amount of evidence showing a negative association between grade retention and educational outcomes. Holmes (1989), in a large meta-analysis, finds that, on average, later test scores of children retained in lower grades are 0.19 to 0.31 standard deviations lower than those of similar children progressing normally through school. The same negative results are reported in a subsequent meta-analysis by Jimerson (2001). There is also a large amount of evidence of a negative relationship between retention and high school (*i.e.* upper secondary) dropout (e.g. Grissom and Shepard, 1989; Roderick, 1994; Jimerson, 1999). The challenge of this literature is of course that grade retention and educational outcomes are

[1] Defined as the share of pupils aged 15 attending a below-reference grade.

likely to be simultaneously determined, which often compromises the identification of a *causal* effect.

There are a few studies providing quasi-experimental evidence on the effects of grade retention. Eide and Showalter (2001) use the variation in the age of entry into kindergarten across US states as an instrument for retention. They find that for white students, grade retention may have some benefit by both lowering dropout rates and raising labour market earnings, although the IV estimates tend to be statistically indistinguishable from zero. Three studies (Jacob and Lefgren (2004, 2009), Roderick and Nagaoka (2005)) exploit a discontinuity in the retention decision under Chicago's high-stakes testing policy[2] introduced in 1996-97. The policy created a discontinuity in the relation between scores in a single standardised test (thereby the label "high stakes") and the probability of grade retention. Using a regression discontinuity design, these studies evaluate the effects of grade retention on pupil performance at different points in time. Jacob and Lefgren (2004) find no systematic differences in performance between retained and promoted students in the short-run. Roderick and Nagaoka (2005) show that third-grade students who were retained do not yield higher language test scores two years after the retention, and that retained sixth graders had lower achievement growth. Finally, Jacob and Lefgren (2009) find that grade retention leads to a modest increase in the probability of dropping out for older students, but has no significant effect on younger students. Finally, Manacorda (2008) exploits a discontinuity induced by a rule in Uruguay Junior High School establishing automatic grade retention for students missing more than 25 days and shows that grade retention leads to a substantial increase in drop-out and lower educational attainment even 4 or 5 years later.

In this paper we exploit a reform in the French-Speaking Community of Belgium in 2001 that (re)introduced the possibility of grade retention at the end of both grade 7 and grade 8. Before then[3], grade retention was not allowed at the end of grade 7. The reintroduction of grade retention in 2001 provides a "natural experiment" to evaluate the effects of grade retention. We use information from the PISA study, measuring performance in a standardised test across OECD countries in Maths, Reading and Science at the age of 15. Pupils who have not

[2] In the mid-1990s, the Chicago Public Schools declared an end to social promotion (i.e. no grade repetition sanctions) and instituted promotional requirements based on standardised test scores.

[3] More precisely in the period 1995-2001

repeated a year should then be in grade 10, thus three grades further than the one affected by the reform. We are able to compare results before the reform (PISA 2000 and 2003) and after the reform (PISA 2006), which is a major advantage in comparison to existing studies. This enables us to compare two different "regimes", with and without grade retention.

We first find that the 2001 decision did lead to a statistically significant change in how 15-year-olds are assigned to grade. The reform led to a reduction of the likelihood of reaching grade 10 at the age of 15 (*i.e.* no grade retention record), and symmetrically, to an increase in the likelihood of attending lower grades (*i.e.* grade 9, 8 or 7).

Compared with many studies, ours also present the advantage of assessing the medium-term effects of grade retention. The reform we examine has (exogenously) changed the likelihood of grade repetition in grade 7 at the age of 12, and we examine the effect of this reform when students are aged 15. However, since these pupils are still below the compulsory school age, we cannot assess the effects of the reform on the *final* educational achievements. Comparing same-age (retained *vs* promoted) pupils in the medium run remains problematic, because they are by definition in different stages of the curriculum.[4]

However, we can nicely test for one necessary condition for grade retention to be justifiable, which is that it should at least succeed in filtering out weaker students from passing to higher grades. That is, in order to provide any grounds to grade retention, one should at least be able to show that, at grade 10, the distribution of score under a "grade retention regime" is better than under a "social promotion regime". [5] We will show that we find supporting evidence for a filtering out effect of the reform.

The data also allow us to compare the attainment of those filtered out under the "grade repetition" regime *vs.* the "social promotion" regime. This allow us to shed some light on two conflicting trends impacting grade repeaters: *i)* a (negative) curriculum effect as repeating a grade means being exposed to a poorer/less demanding curriculum than the one taught in the (higher) reference grade[6]; and *ii)* a lower-ability/less-demanding curriculum (positive)

[4] They attend different grades, as can be seen in Table 2 for instance.

[5] Synonymous with no grade-repetition sanctions.

[6] Grade 10 in Belgium at the age of 15.

matching effect. The latter effect directly echoes the argument of the proponents of grade repetition: weaker pupils should benefit from being exposed longer to a simpler curriculum that better matches their ability and/or attainment.

As to the methodology used in this paper, it is important to stress that the main results are based on the synthetic control (SC) method (Abadie, Diamond, Hainmueller, 2007), which uses data-driven procedures to construct an adequate comparison group/counterfactual. In practice, it is difficult to find a single unexposed unit (here an educational system) that approximates the most relevant characteristics of the French-Speaking Community of Belgium's education system and would provide a counterfactual. The idea behind the synthetic control approach implemented here is that a *combination of countries — a synthetic control* — offers a better comparison than any single country/entity alone (say the Flemish-Speaking Community of Belgium, France, Germany or the Netherlands).

The remainder of this paper is organised as follows. Section 2 is introductory and mainly consists of stylised facts. It documents the international evidence on retention rates and overall PISA scores. It essentially shows that there is *no correlation* between cross-country variance in grade assignment of 15 years-olds and (1) their average score and (2) the dispersion of their scores. Section 3 presents the 2001 reform in the French-Speaking Community of Belgium and documents its impact on the incidence of grade retention using both administrative data and various waves of the PISA survey. It then examines the relationship between (more) grade retention and PISA scores in the French-Speaking Community of Belgium, using the SC method to generate the best possible counterfactual. The plausibility of a filtering out assumption is examined first. Second, the paper looks at how the score of filtered out students compares under the two regimes. Section 4 concludes.

2. Grade assignment and grade retention: the international evidence

The different OECD countries that participated to the three waves of PISA (2000, 2003, 2006) provide a relatively large source of variance as to the incidence of grade retention (see Annex 1 to 3). Using country-level aggregate data, it is easy, in Figure 1, to see how the share of pupils attending the grade of reference[7] (our proxy for the intensity of grade retention)[8] relates to score. Figure 1 basically suggests an absence of correlation between the importance of grade retention (*i.e.* a leftward shift) and average score in math. Similar results are obtained for reading and science scores. Note incidentally that Figure 2 conveys the same information about the relationship between grade retention and standard deviation of PISA scores.

However, country-specific unknown factors may be systematically correlated with *i)* the (varying) propensity of countries to resort to grade retention and *ii)* scores. Under these circumstances the results of an analysis exploiting the inter-country variance are bound to be biased.

This is why it is worth focusing solely on the intra-(or *within-)* country variance. This is made possible by the availability of three consecutive waves of the PISA survey (2000, 2003 and 2006). Exploiting the country-level panel structure of PISA is thus possible to re-examine the relationship we are interested in. Descriptive results are displayed in Figure 3 and Figure 4. They tend to confirm the absence of relationship between the (within country) evolution of score from 2000 to 2006 and the changing proportion of pupils who attend Grade 10 at the age of 15.

The descriptive results on display in Figures 4 & 5 are confirmed by the OLS estimation of equation [1] (Table 1). The latter uses the same data aggregated at country level. It includes country fixed effects to retain the within-country part of the variance. The list of controls includes a year trend -- that captures changes that are common to the whole group of countries sampled -- and a vector of socio-economic background variables (Table 1). Note finally that this model is estimated separately for each of the topics covered by PISA (Math, Science and Reading literacy).

[7] Grade 10 in most countries, grade 9 otherwise. The grade of reference is identified as the most attended one among 15 year-olds who participated to PISA.

[8] Of course, differences could also be due to differences in entry school ages. In the case of Belgium, except in rare exceptions, pupils enter grade 1 during the calendar year they turn 6. The exact cut-off date is the 1st of January. All the pupils that have reached the age of 6 before that date must start grade 1 during the calendar year that ends on the 1st of January.

$Y_{i,t}= \alpha + \beta SREFG_{i,t} + Z'_{i,t}\gamma + \delta YEAR + \eta_i + \varepsilon_{i,t}$ [1.]

i= 1,…., J and t=2000, 2003, 2006

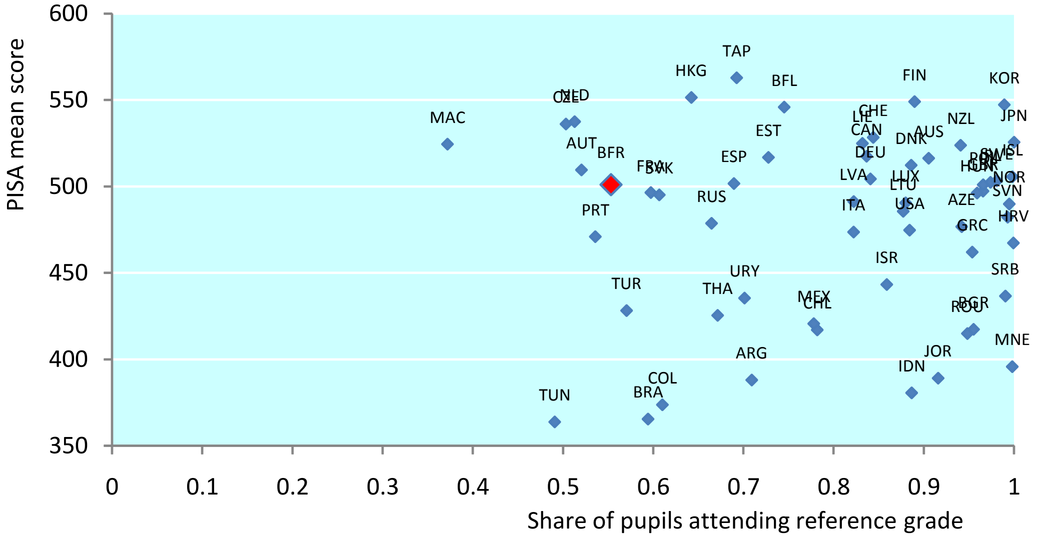
where

- $Y_{i,t}$ is the average PISA score of country *i* during year *t*;
- $SREFG_{i,t}$ is the share of pupils attending the reference grade[9] in country *i* during year *t*;
- $YEAR_t$ is the year of observation capturing a trend that would be common to all countries ;
- $Z'_{i,t}$ is a vector of controls that include the average parental socio-economic background index and education attainment;
- η_i is the country *i* fixed effect ;
- and $\varepsilon_{i,t}$ a random error term centred on zero ;

A major limitation however is that a *within* (country) restriction, as we imposed in the previous section (Figure 3 & 4 or equation [1]), could prove insufficient to properly identify the effect on scores of the grade-assignment regime. Indeed, changes observed within a country over time may be driven by unobserved confounding factors that are correlated with scores, like a better economic environment (insufficiently or inadequately captured by the observables *Z)*. Thus, ideally the identification of the effects of grade retention requires not only an exogenous change in grade repetition, but also the existence of a counterfactual for comparison. This is why we now propose an analysis comparing the changes observed in the French Community to the changes observed in a control group.

[9] Grade 10 in the French-Speaking Community of Belgium, like in most of the other countries considered here.

Figure 1 – Average score in math and share of pupils aged 15 attending reference grade[a]. Year 2006

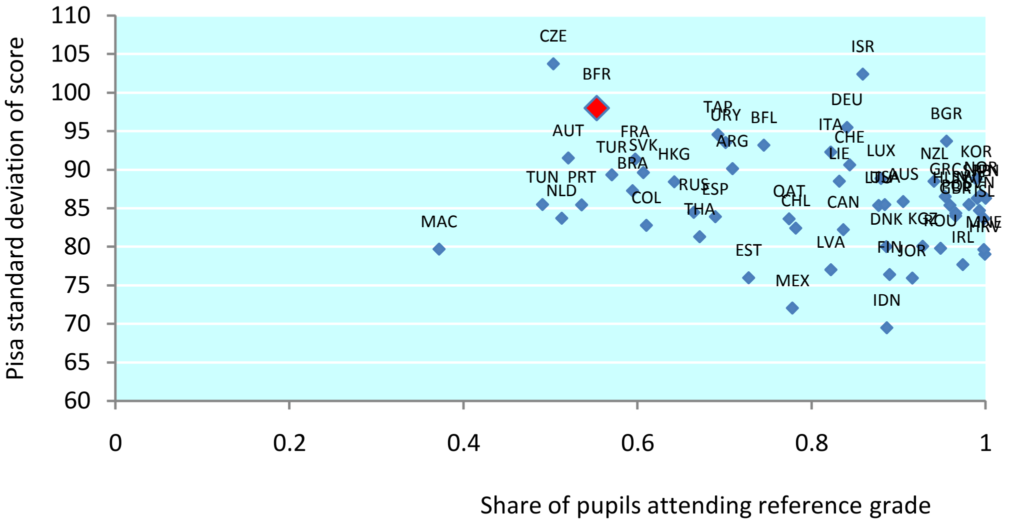


a) Grade 10 in most countries, grade 9 otherwise. The grade of reference is identified as the most attended grade among 15 year-olds who participated to PISA.

ARG : Argentina ; AUS : Australia ; AUT : Austria ; AZE : Azerbaijan ; BFR : French-Speaking Community of Belgium; BFL: Flemish-Speaking Community of Belgium; BGR : Bulgaria ; BRA : Brazil; CAN : Canada; CHE : Switzerland; CHL : Chile COL : Colombia CZE : Czech Republic; DEU : Germany; DNK : Denmark; ESP : Spain EST : Estonia; FIN : Finland; FRA : France; GBR : United Kingdom; GRC : Greece; HKG : Hong Kong-China; HRV : Croatia; HUN : Hungary; IDN : Indonesia IRL : Ireland; ISL : Iceland; ISR : Israel; ITA : Italy JOR : Jordan; JPN : Japan KGZ : Kyrgyzstan; KOR : Korea LIE : Liechtenstein LTU : Lithuania LUX : Luxembourg; LVA : Latvia; MAC : Macao-China; MEX : Mexico; MNE : Montenegro; NLD : Netherlands; NOR : Norway; NZL : New Zealand; POL : Poland; PRT : Portugal QAT : Qatar; ROU : Romania; RUS : Russian Federation; SRB : Serbia; SVK : Slovak Republic; SVN : Slovenia SWE : Sweden; TAP : Chinese Taipei; THA : Thailand TUN : Tunisia; TUR : Turkey; URY : Uruguay; USA : United States.

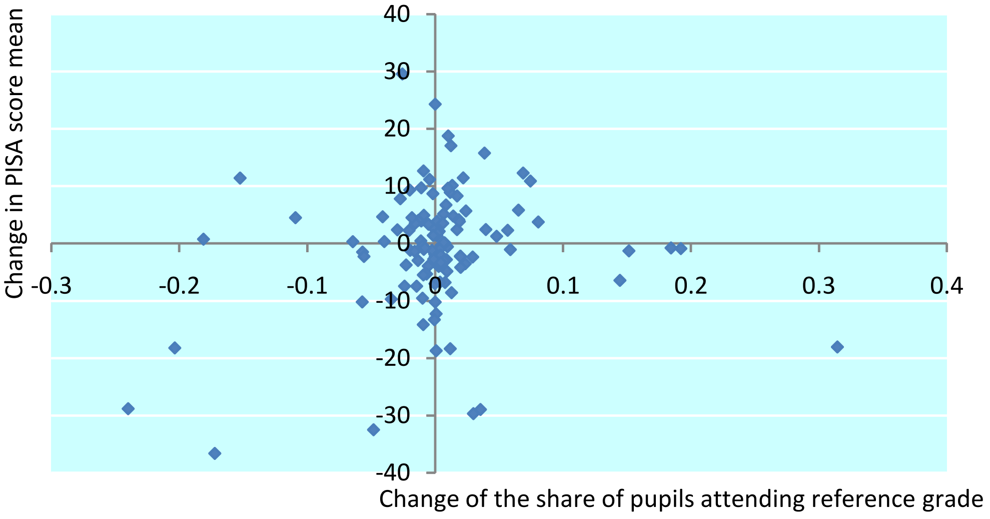
Source: PISA 2006

Figure 2 – Standard deviation of score in math and share of pupils aged 15 attending reference grade[a]. Year 2006



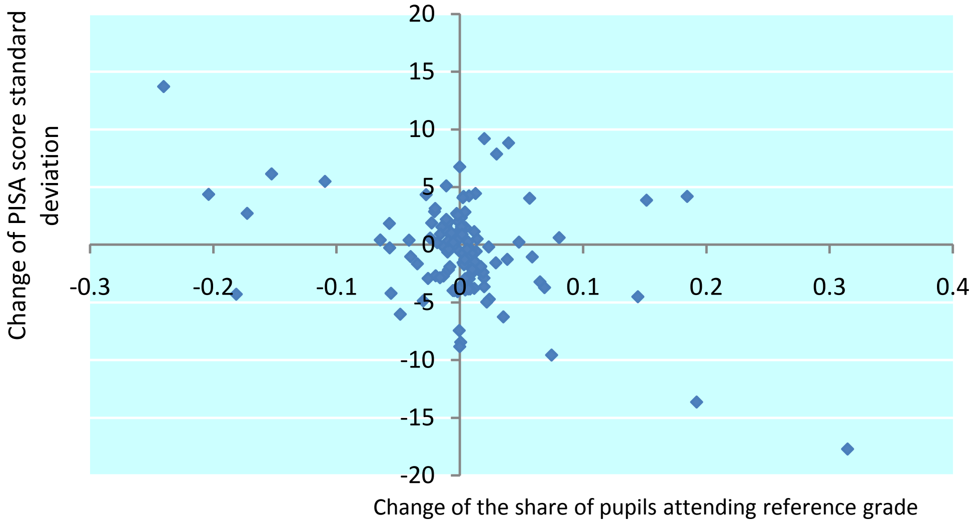
a) Grade 10 in most countries, grade 9 otherwise. The grade of reference is identified as the most attended grade among 15 year-olds who participated to PISA.
Source: PISA 2006

Figure 3 – Within country change of the share of age 15 pupils attending reference grade and change of average score. Math.



Source: PISA 2000, 2003 and 2006

Figure 4 – Within country (statistically significant) change of the share of age 15 pupils attending reference grade and change of standard deviation of score. Math.



Source: PISA 2000, 2003 and 2006

Table 1 – Shares of pupils in reference grade and PISA scores (OLS coefficients). Within analysis.

	Country average score			Country standard deviation		
Variable	**math**	**read**	**scie**	**math**	**read**	**scie**
Share of pupils attending reference grade[a]	0.15	-0.10	-0.12	-0.23	-0.09	-0.16
	(-0.460)	(0.792)	(0.575)	(0.003)	(0.469)	(0.038)
R^2	0.97	0.86	0.92	0.78	0.62	0.74
N obs	132	129	132	132	129	132

P-values are between brackets Controls include country fixed effects, year, average highest parental socio-economic index, average highest degree of father, average highest degree of mother.

a) Grade 10 in most countries, grade 9 otherwise. The grade of reference is identified as the most attended grade among 15 year-olds who participated to PISA.

Source: PISA 2000, 2003 and 2006

3. Exploiting the French-Speaking Community reform

3. 1. The 2001 reform in the French-Speaking Community of Belgium: a source of exogenous variation of grade retention

Grade retention/retention and different-grade assignment of same-age pupils have existed for a long time in Belgium, and is particularly frequent in the French-Speaking Community (Figure 1).[10] The retention decision is based on the teachers' assessment of the pupil's ability of passing to a higher grade. There is no standardised test used across schools, nor is there a clearly defined threshold to determine whether a pupil should be retained or not. All pupils do take exams at the end of the school year, for each subject, and the retention decision is made after these exams have been taken.

Opponents to grade retention succeeded in 1995 in suppressing grade retention at the end of grade 7 (1st year of secondary education). From 1995 to 2001 no grade retention was allowed at the end of grade 7 (1st year of secondary education), a decision that translated into a sharp fall in the number of "repeaters" (Figure 5). During that period, grade retention sanctions could only be pronounced at the end of grade 8. Pupils could only possibly repeat grade 7 upon agreement between parents and teachers. This is why on Figure 5 one observes a persistence of grade retention at the end of grade 7 during the 1995-2001 period.

The proponents of grade retention made a successful comeback in September 2001, when the decision was taken[11] to re-establish the possibility to retain weak students in grade 7. In a few words, the 2001 reform was such that after the school year 2001-02 it became possible to repeat grade 7 or grade 8, although not both.[12] Administrative data (Figure 5) show that the

[10] Belgium is a federal state where the educational policy is split according to linguistic lines. Each linguistic community is in charge of its educational system. Only minor aspects of the educational policy (like the age of compulsory education) remain under federal jurisdiction

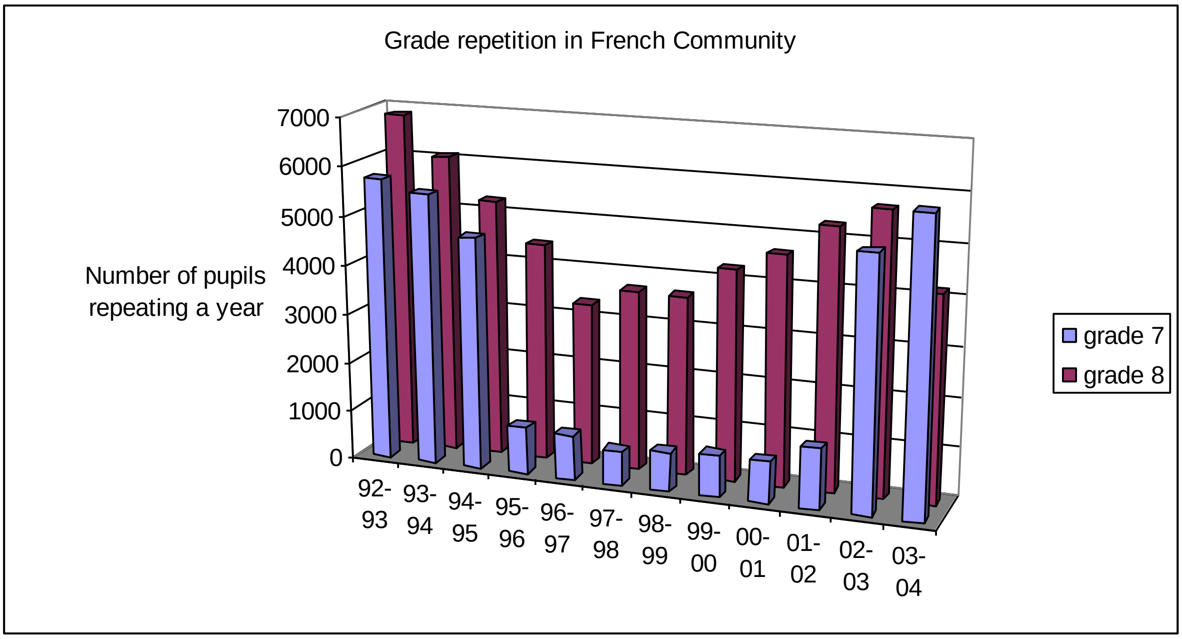
[11] *Décret relatif à l'organisation du premier degré de l'enseignement secondaire* D. 19-07-2001 M.B. 23-08-2001

[12] Formally, the legislator insists on the fact that the reform's aim was not exactly to force the pupils to "repeat" the year, but to channel weaker students (who did not achieve satisfactory results at the end of grade 7 or at the end of grade 8) towards a "complementary" year. In practice, however, it amounts to imposing that these students take more time before moving to the upper grade.

number of pupils repeating grade 7 consequently rose sharply from the school year 2002-03 onwards. The same data also show that the total number of students repeating grade 7 or grade 8 is substantially higher after 2001, meaning that the 2001 reform generated an overall increase of the risk of being retained into a lower grade.

Thus, the 2001 reform enables us to compare a system with grade retention with a system with (almost) no grade retention in grade 7. Hereafter, we exploit the 2001 reform and investigate the medium-term[13] (causal) effects of the reform on the PISA scores.

Figure 5 – Incidence of grade retention at Grade7 and Grade 8. School year 1992-93 to 2003-04



Source: French-Speaking Community of Belgium, Ministry of Education.

[13] Remember that we look at age 15 scores to identify the effect of a decision that affected pupils when they were aged 12-13.

3.2. Using the Synthetic Control Method to generate a counterfactual

To assess the effects of the reform, we use a *synthetic country* (SC) as a control (Abadie, Diamond, Hainmueller, 2007). The method generalizes the commonly used difference- in-difference model. The SC method a priori uses all countries other than French-Speaking Belgium that participated in PISA as potential controls. The key idea is to identify a linear combination of the other $i=2$ to J countries — $W=(w_{2,}....w_J)$ such that $w_i \geq 0$ and $w_2+....+w_J= 1$ — that best reproduces the French-Speaking Community of Belgium (*i.e.* the treated entity) during the pre-reform period (*i.e.* 2000 and 2003), both in terms of average attainment Y and a list of observed controls Z that potentially affect attainment.[14] The identification of the effect of the reform is achieved by comparing the post-reform *observed* average attainment of pupils in *i)* the French-Speaking Community of Belgium Y_1 and *ii)* its synthetic equivalent $\mathrm{Y}_{SC}=.\sum w_i Y_i$, $i=2$ to J.

Annex 1 explains how this is done analytically and why the post-treatment (*i.e.* 2006) first difference between the treated and the synthetic control entities properly identifies the effect of treatment in the presence of unobserved time effects and country effects that are not randomly distributed.

i) PISA Evidence of more grade retention as a consequence of the 2001 reform

Before turning to the implementation of this evaluation strategy, we need to complement the information highlighted in Figure 5 and check that the PISA data used here also contain robust evidence that the reform has generated some change in the French-Speaking community of Belgium in the likelihood of experiencing grade repetition.

Table 2 reports the distribution of pupils aged 15 according to their grade in French-Speaking Belgium and in the synthetic control entity. We see that in the French-Speaking Community

[14] Our list of controls/predictors include student/teacher ratio, ratio of computers to school size, % of teachers with proper certification, mother education, father education, the highest parental socio-economic index (HISEI), and the share of pupils attending the reference grade prior to the reform.

of Belgium, *less* pupils aged 15 reached grade 10 in 2006 (*i.e.* after the reform) than in 2003 or 2000 (before the 2001 reform), and, symmetrically, that more pupils were below grade 10.

Frequencies reported in Table 2 are direct sign that more grade retention (with lasting effects) occurred in the French-Speaking Community of Belgium, as the only way to be at age 15 in grade 10 is to have a no-grade-retention record. In short, all this accords with the grade-retention regime change introduced in the French-Speaking Community of Belgium in Sept. 2001.

Table 2 – Share of pupils aged 15 attending grade 10 *vs.* grade < 10 (%) in the French-Speaking Community of Belgium

	French-Speaking Community of Belgium	
2000	0.59	β_{00}
2003	0.59	β_{03}
2006	0.55	β_{06}

Source: Pisa 2000, 2003 and 2006

ii) The screening-out test

Grade retention to be justifiable should at least succeed in filtering out weaker students from passing to higher grades. That is, in order to provide any grounds to grade retention, one should at least be able to show that, conditional on grade, the distribution of scores under a "grade retention regime" is on average better than the distribution of scores under a "social promotion regime".[15]

We will focus here on Grade 10 and use the SC method to generate the counterfactual. Estimated weights, for each of the models estimated here using the SC method, are reported in Annex 9, first Table.

[15] Those who make it to grade 10 for instance should be, *ceteris paribus*, better than under the less selective regime.

Assume that they are (potentially) two categories of students attending grade 10. First, the non-delayed students, unaffected[16] by the grade-repetition regime change. These always attend grade10 at the age of 15. Let us note their average score YND^{10}. The second group consists of the $0<\mu<1$ students directly affected by the reform (the "treated students" hereafter). Their average score in 2006 (in grade10) is YT^{10}.

Assume further that Y_1^{10} is the post-reform *observed score average*[17] *of grade 10* students from the French-Speaking Community of Belgium. It is solely driven by the attainment of non-delayed students.

$$Y_1^{10} \equiv YND^{10} \qquad [2.]$$

By comparison, the grade 10 score average in the synthetic French-Speaking Belgium — estimated using the data-driven SC procedure exposed above — should be a linear combination of *i)* the score of non-delayed/non-treated students and *ii)* that of "treated" students (those who reached grade10 at the age of 15 thanks to a less stringent grade-repetition regime before the implementation of the 2001 reform). Note that π — the fraction of the cohort that has been "treated" normalised by the size of grade10 in 2006 (β_{06}) to properly capture the weight of the treated in the 2006 grade 10 average — can be estimated using Table 2 figures [(58.68-55.32)/(55.32)=0.0608]. It is about 6.8 %

$$Y_{SC}^{10} \equiv (1-\pi)YND^{10} + \pi YT^{10} \qquad [3.]$$

where $\pi = \mu/\beta_{06}$ and $\mu = \beta_{03}-\beta_{06}$ as reported in Table 2

Now, turning to grade 10 score average differences, using [2] and [3] we get:

$$Y_1^{10} - Y_{SC}^{10} = YND^{10} - (1-\pi)YND^{10} - \pi\, YT^{10} \qquad [4.]$$

[16] We leave aside spillover effects.

[17] All results presented hereafter use students' average scores (*i.e.* the unweighted average of their math, science and reading PISA scores).

or equivalently,

$$Y_1^{10} - Y_{SC}^{10} = \pi(YND^{10} - YT^{10}) \qquad [5.]$$

where $\pi = \mu / \beta_{06}$

The estimation of the left-hand part of expression [5] gives a direct indication of the score gap between the treated and the non-treated students. If $Y_1^{10} - Y_{SC}^{10}$ is significantly positive then one can conclude that score of the treated/"socially promoted" students was below those who usually attend grade 10 at the age of 15. In that case, it can be inferred that the reform properly managed to filter out weaker students (those who presumably could benefit from a less demanding curriculum or extra time).

An important restriction is that [5] provides an estimate of the *π-weighted* relative performance of treated students within grade 10. Hence, it is likely that the 1.71 gap reported on the first line of Table 3 *underestimates* the actual score gap. A short-cut strategy to cope with this bias consists of "dividing" $Y_1^{10} - Y_{SC}^{10}$ by $\pi = 0.061$. This rapid transformation suggests a positive score gap of 28 points.[18]

But the 1.71 points estimate on the first line of Table 4 is so uncertain (*i.e.* p-value= 0.40) that there could be no effect at all, or even a positive one. Further econometric analysis is highly desirable to test the plausibility of that result.

Our strategy in that respect is simple. It consists of incrementally eliminating the upper percentiles of the grade 10 distribution of scores (Figure 6) to increase the (relative) weight of treated students in the comparison. The crucial assumption is that treated students must be concentrated just at the bottom of the grade 10 distribution.

Formally, we estimate:

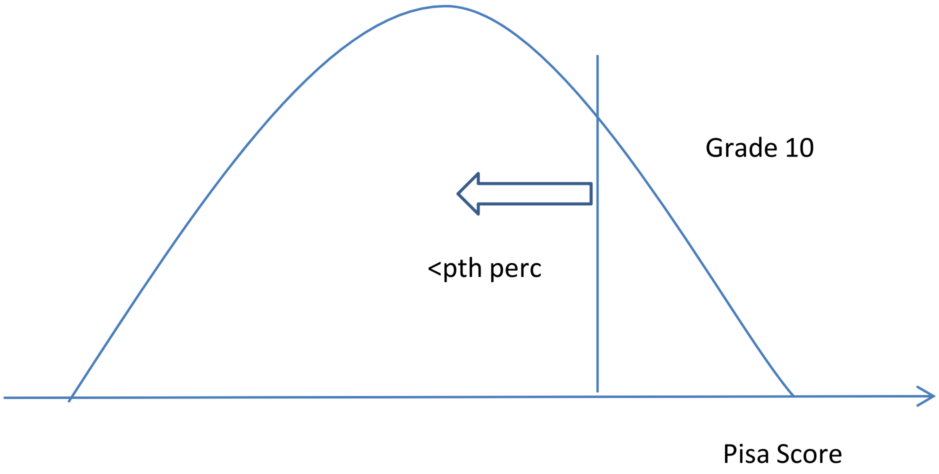
$$Y_1^{10} - Y_{SC}^{10} \,\big|\, (y^{10} \leq \theta^{th}\ perc.) = \pi(\theta)(YT^{<10} - YT^{10}) \qquad [6.]$$

[18] Pisa scores have an (international) average of 500 and a standard deviation of 100.

- θ ranging from 90 to 10.
- $\pi(\theta) = \mu / \theta\beta_{06}$ assuming all treated pupils belong to the retained percentiles

Results are reported in Table 3. They contain statistically significant evidence[19] that $YND^{10} > YT^{10}$ and thus that more grade repetition after the Sept. 2001 reform has primarily led to the retention of students who had PISA scores inferior to the grade 10 average.

Figure 6 – Increasing the chances of identifying the sorted-out students: eliminating the upper end of the Grade 10 score distribution.[20]



[19] See Annex 5 for a presentation of inference analysis/hypothesis testing with SC.

[20] The actual score distribution is reported in Annex 6.

Table 3 – Grade 10 change of attainment between French-Speaking Belgium and synthetic French-Speaking Belgium as an estimate of the treated vs non-treated student gap. Year 2006$^{\$}$

SC_equ	_Y_French Speaking Belgium	_Y_synthetic	Y_dif	Probt	π	θ	$\pi(\theta)=\mu/\theta$	Y_diff/$\pi(\theta)$
All obs	541.66	539.93	1.73	0.4084	0.061	1.00	0.061	28.48
1<=p90	528.76	528.02	0.74	0.5431		0.90	0.068	10.98
1<=p80	517.50	507.20	10.30***	0.0000		0.80	0.076	135.41
1<=p70	506.19	502.26	3.93**	0.0200		0.70	0.087	45.21
1<=p60	494.15	486.87	7.28***	0.0001		0.60	0.101	71.79
1<=p50	481.38	477.18	4.20**	0.0132		0.50	0.122	34.52
1<=p40	466.97	456.29	10.68**	0.0000		0.40	0.152	70.21
1<=p30	449.92	438.77	11.15***	0.0000		0.30	0.203	54.97
1<=p20	427.71	422.00	5.70**	0.0127		0.20	0.304	18.74
1<=p10	394.68	397.97	-3.29	0.3477		0.10	0.608	-5.40

Source: PISA 2000, 2003 and 2006

*** Significant at 1%, **significant at 5%, * significant at 1%

$^{\$}$ Score comparisons for the year 2000 to 2006 for a selection of estimated models are on display in Appendix 7, whereas Annex 8 displays the comparison of predictor/control variables.

iii) How do filtered-out students fare?

Do the filtered out students achieve better/worse when they repeat (and attend a lower grade) than when they are "socially promoted" (and attend the reference grade 10)? An answer to that question can be provided by comparing French-Speaking Belgium's overall (*i.e.* all grades pooled) score average to its synthetic counterfactual.[21]

Assume now that they are three categories of students forming the public of both grade 10 and grade<10. First, the non-delayed students, unaffected[22] by the grade repetition regime change. These always attend grade10 at the age of 15. Keep noting their average score YND^{10}. Another group — also unaffected by the regime change — consists of the delayed students always attending grade <10. Their score is noted $YD^{<10}$. The third group consists of the μ students directly affected by the reform (again, the "treated students"). Their average

[21] The computed weights used to build synthetic controls are presented in Annex 9.

[22] Again, we leave aside spillover effects.

score in 2006 (in grade10) is YT^{10} . Whereas their score in grade<10 in 2006 after the reform is $YT^{<10}$.

Assume further that Y_1 represents the post-treatment (i.e. 2006) *observed* overall score average in the French-Speaking Community of Belgium.

$$Y_1 \equiv \alpha_0 YND^{10} + (1\text{-}\alpha)YD^{<10} + \mu YT^{<10} \qquad [7.]$$

with $\alpha_{0+}\,\mu = \alpha$

The synthetic counterfactual — corresponding to the case where, due to the absence of the equivalent of the 2001 reform, μ students are in grade 10 —, writes;

$$\mathrm{Y}^1_{SC} \equiv \alpha_0 YND^{10} + \mu YT^{10} + (1\text{-}\alpha)YD^{<10} \qquad [8.]$$

with $\alpha_0 + \mu = \alpha$

The difference between these two observed averages [7],[8] is equal to

$$Y^1 \text{-} Y^1_{SC} = \mu(YT^{<10} \text{-} YT^{10}) \qquad [9.]$$

with $0<\mu<<<1$.

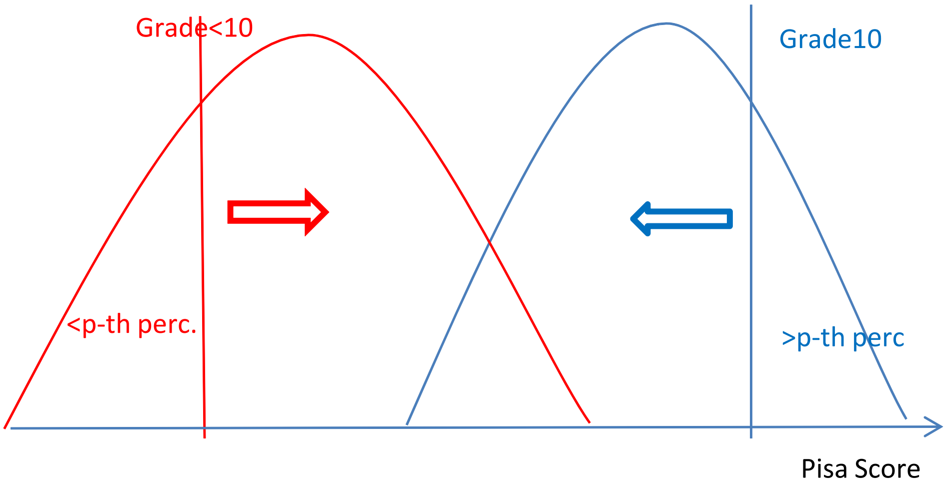
The first line of Table 4 reports estimates of [9]. They suggest a minor attainment decrease (*Y_dif*) of about -0.29 points which appears totally insignificant from a statistical point of view. But again, these results consist of averages that are computed with the scores of all pupils. They implicitly (and wrongly) assume that all pupils have been "treated".

Turning back to the frequencies of Table 2, it is more likely that only a small fraction of the cohort that has been directly[23] “treated” ($\mu = \beta_0 - \beta_1$=0.586 -0.553= 0.0336): about 3.4%. Hence, average-based comparisons of the kind reported in Table 3 — and also in previous sections — are unlikely to properly reveal the true magnitude of treatment on those who have really been treated.

[23] We leave aside spillover effects.

To cope with this problem, we follow a strategy that is similar to the one used just above. It consists of incrementally eliminating the upper (and lower) percentiles of the distribution of scores within each grade (Figure 7)[24] to lift the (relative) weight of treated students in the averages that are compared with the SC method. The crucial assumption is now that the "treated" students must be concentrated above and below the grade<10/grade10 cut-off zone.

Figure 7 – Increasing the chances of identifying the treated students: eliminating upper and lower parts of the grade-specific score distribution



If for instance one eliminates from the SC computation the students that are above (below) the 90-th (10-th) percentile of grade 10 (grade<10) (meaning that we retain θ=90% of the initial overall sample**)**, we should *a priori* increase the weight of the treated students in the comparison of averages.

$$Y_1 - Y_{SC}\left|(y^{<10} > 10^{th}\ perc.\ or\ y^{10} \leq 90^{th}\ perc.) = \mu(0.9)(YT^{<10} - YT^{10})\right. \qquad [10.]$$

where $\mu(0.9) = \mu * 1/0.9$ assuming all treated pupils belong to the kept percentiles

The second line of Table 4 contains the results when one estimates the left-hand part of [9]. They suggest a -1.02 (non-weighted) effect that is not statistically significant. When divided

[24] Actual score distribution by Grade in Annexe 5

by the weighing factor $\mu(0.9)$ the estimate is -27 points, which suggest that the reform has led to lower scores for the treated students. But again this result is not statistically significant.

The rest of Table 4 presents our estimates when one eliminates 20% ,30%, 40% … up to 90% of the initial sample to focus on the observations concentrated below and above the cut-off point where, presumably, treated students should be concentrated.

$$Y_1 - Y_{SC} \left|(y^{<10} > 100 - \theta^{th} \text{ perc. or } y^{10} \leq \theta^{th} \text{ perc.}) = \mu(\theta)(YT^{<10} - YT^{10})\right. \qquad [11.]$$

where $\mu(\theta) = \mu * 1/\theta$ assuming all treated pupils belong to the retained percentiles

Basically, results remain unchanged. The sign of the estimates change from negative to positive. What is more, all estimates are statistically non significant. The tentative conclusion is thus that the 2001 reform has had no effect on the score of filtered-out students.

This is not necessarily surprising. Recall there are two opposite mechanisms which could affect the score of students when they are moved from grade 10 to grade<10 (or vice versa): First, a (negative) curriculum effect implying that being the grade<10 curriculum is poorer than the one taught in grade 10. Second, a (positive) ability/curriculum matching effect implying that the weakest students attending grade 10 before the 2001 reform could be those who struggle to grasp the material of a more advanced curriculum, thereby benefiting from being retained in grade <10 where there are exposed to a curriculum that better matches their capabilities.

Since the reform of 2001 lead to reallocation of pupils from grade 10 to below, one would expect a negative difference in means due to a curriculum effect. But our results suggest that this has probably been compensated by improvements due to a better ability/curriculum match.

Table 4 – All grades pooled. Change of attainment between French-Speaking Belgium and synthetic French-Speaking Belgium as an estimate of treated students attainment change. Year 2006[$]

Equation	**_Y_French Speaking Belgium**	**_Y_ synthetic**	**Y_dif**	**p-value**	***μ***	***θ***	***μ(θ)=μ/θ***	**Y_diff/μ(θ)**
All obs.	493.40	493.70	-0.29	0.8537	0.034	1	0.034	-8.67
>p10<=p90	499.97	500.98	-1.02	0.5056		0.9	0.037	-27.18
>p20<=p80	500.48	497.94	2.54	0.0915		0.8	0.042	60.31
>p30<=p70	500.59	502.29	-1.70	0.3039		0.7	0.048	-35.33
>p40<=p60	500.42	503.02	-2.59	0.1583		0.6	0.056	-46.18
>p50<=p50	500.61	503.51	-2.90	0.1692		0.5	0.067	-43.01
>p60<=p40	501.34	503.09	-1.75	0.5013		0.4	0.084	-20.81
>p70<=p30	503.03	504.21	-1.18	0.7101		0.3	0.112	-10.52
>p80<=p20	507.91	504.32	3.59	0.3660		0.2	0.168	21.34
>p90<=p10	526.04	515.72	10.32	0.1533		0.1	0.337	30.65

Source: PISA 2000, 2003 and 2006
*** Significant at 1%, **significant at 5%, * significant at 1%
[$] Score comparisons for the year 2000 to 2006 for a selection of estimated models are on display in Appendix 7, whereas Annex 8 displays the comparison of predictor/control variables.

4. Conclusion

This paper exploits a reform in the French-Speaking Community of Belgium (re)introducing the possibility to impose grade retention at the end of both grade 7 and grade 8, to evaluate the effects of grade retention. The reform constitutes a "natural experiment" introducing an exogenous variation in the assignment of pupils to grade. Indeed, the reform lead to a reduction in the likelihood of reaching grade 10 at the age of 15 (*i.e.* no grade retention record), and symmetrically, to an increase in the likelihood of attending lower grades.

Using a synthetic control (SC) method to generate a post-reform French-Speaking-Belgium counterfactual we are able to address two issues. First, whether a grade retention regime does at least succeed in filtering out weaker students. Second, whether the weaker pupils who end up being retained into lower grades under a "grade repetition regime" perform worse/better than under a "social promotion" regime. We find statistically significant evidence in support of the screening out effect of grade retention. But we fail to demonstrate that filtered out students perform differently under the “grade repetition” regime than under the “social promotion” regime. Our results suggest that the negative curriculum effect repeaters traditionally suffer from may have been compensated by a better ability/curriculum match.

A limitation of the paper — that uses same-age score data — is that it cannot assess the effects of the reform on the *final* educational achievements. Comparing retained and promoted pupils at the age of 15 is problematic, as, by definition, they are in different stages of the curriculum.

In particular those who are forced to repeat a grade and who suffer from a negative curriculum effect should normally benefit from a richer curriculum when — eventually — they get promoted to the higher grade. The long-run balance could then perhaps be that grade repetition has a positive effect. However, the proper long-run cost-benefit analysis of grade repetition should then also account for the large costs of grade retention, particularly in terms of opportunity costs for the pupils (each grade repetition means that one year is lost), but also in terms of teaching resources (each grade repetition means one extra year of funding).

Bibliography

Abadie, A., A. Diamond and J. Hainmueller (2007), Synthetic Control Methods For Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program, *NBER Working Paper*, No. 12831, Ma.

Eide, E.R. and M.H. Schowlater (2001), The Effect of Grade Retention on Educational and Labour Market Outcomes, *Economics of Education Review*, 20(6), pp. 563-576.

Grissom, J. and L. Shepard (1989), *Repeating and dropping out of school*, in L. Shepard and M. Smith (Eds.), Flunking Grades: Research and policies on Retention, London: The Palmer Press.

Holmes, C. T. (1989), *Grade level retention effects: A meta analysis of research studies*. In L. A. Shepard & M. L. Smith (Eds.), Flunking grades: Research and policies on retention (pp. 16-33). London: Falmer.

Jacob, B.A & L. Lefgren, (2004), Remedial Education and Student Achievement: A Regression-Discontinuity Analysis, *Review of Economics and Statistics*, LXXXVI(1), pp. 226-244.

Jacob, B.A & L. Lefgren, (2009), The Effect of Grade Retention on High School Completion, forthcoming in *American Economic Journal - Applied Economics*.

Jimerson, S.R. (1999), On the Failure of Failure, *Journal of School Psychology*,37(3), pp. 243–272.

Jimerson, S.R. (2001), Meta-analysis of Grade Retention Research: Implications for Practice in the 21st Century, *School Psychology Review* 30(3), pp. 420-437

Manacorda, M. (2008), The Cost of Grade Retention, *CEP Discussion Papers* 0878, Centre for Economic Performance, LSE.

McCoy, A.R. & A. J. Reynolds (1998), Grade Retention and School Performance: An Extended Investigation, *Institute for Research on Poverty, Discussion Paper no. 1167-98,* University of Wisconsin,-Madisson.

Roderick, M. (1994), Grade retention and school dropout: Investigating the association. *American Educational Research Journal*, 31, pp. 729–759.

Roderick, M., & Nagaoka, J. (2005). Retention under Chicago’s high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis*, 27(4), pp. 309-340.

Annex 1 – PISA 2000, descriptive statistics

		Mean							Standard deviation						
Country	Nobs	Share of pupils attending ref. grade	Parental SES index	Mother degree	Father degree	Score in Math	Score in reading	Score in Science	Share of pupils attending ref. grade	Parental SES index	Mother degree	Father degree	Score in Math	Score in reading	Score in Science
Australia	2859	0.92	52.64	4.07	4.09	530.33	529.94	523.43	0.26	16.75	1.07	1.07	89.26	99.58	95.68
Austria	2640	0.50	49.02	3.41	3.49	506.86	508.16	512.59	0.50	14.03	0.95	0.96	89.89	109.02	89.37
Belgium((Fl)	2211	0.74	48.33	4.33	4.43	546.16	493.55	524.44	0.44	16.29	1.09	1.03	92.95	101.03	87.83
Belgium (Fr)	1573	0.59	50.67	4.05	4.20	493.60	505.25	452.32	0.49	17.21	1.27	1.20	105.10	78.80	115.24
Brazil	2717	0.41	42.77	3.11	3.13	320.05	549.54	388.10	0.49	17.23	1.40	1.43	93.77	82.66	103.52
Canada	16701	0.80	51.27	4.64	4.48	524.98	524.63	515.59	0.40	16.37	0.86	1.00	79.69	94.46	85.38
Switzerland	5456	0.87	47.73	3.41	3.61	531.36	477.07	503.18	0.34	15.37	1.02	1.10	85.23	92.00	95.52
Czech Republic	3066	0.57	48.32	4.13	3.99	499.27	495.16	494.82	0.50	13.70	1.03	1.03	93.16	108.89	101.84
Germany	2830	0.84	49.75	3.60	3.88	500.03	433.52	497.15	0.37	15.80	1.00	1.04	95.81	98.92	96.26
Denmark	2395	0.92	49.82	4.26	3.98	516.02	446.48	492.27	0.27	15.86	1.03	1.06	80.49	109.92	96.58
Spain	3428	0.72	45.02	3.20	3.36	478.79	521.75	491.11	0.45	16.40	1.38	1.42	84.85	82.15	90.61
Finland	2703	0.89	50.07	3.68	3.48	537.03	405.06	527.14	0.31	16.40	1.19	1.15	74.56	101.55	86.99
France	3861	0.60	47.89	3.94	3.84	523.24	504.78	497.60	0.49	17.79	1.15	1.15	87.23	92.90	93.72
United Kingdom	54627	0.54	44.34	4.06	3.74	519.82	494.52	526.15	0.50	18.51	1.01	1.11	107.19	97.29	110.34
Greece	2605	0.95	47.89	3.84	3.79	446.42	491.02	472.65	0.21	18.07	1.30	1.31	100.67	93.20	91.77
Hungary	3491	0.95	48.57	4.17	3.92	492.52	490.00	506.26	0.22	15.77	1.01	1.01	88.02	86.80	92.92
Ireland	2128	0.96	48.23	4.03	3.72	502.92	510.17	512.89	0.19	15.22	1.23	1.31	78.77	88.76	94.95
Iceland	6424	1.00	54.11	3.83	3.93	526.54	518.18	521.76	0.07	15.38	1.24	1.20	78.24	90.72	80.15
Italy	4413	0.83	46.79	3.84	3.82	460.25	492.97	481.98	0.38	15.49	1.16	1.14	82.82	87.12	94.86
Japan	2940	1.00	50.37			560.07	528.00	522.15	0.00	15.49			81.13	83.42	98.35
Korea	2769	0.99	42.41	3.65	3.85	541.47	505.93	526.14	0.11	14.24	1.24	1.22	79.79	93.79	87.89
Liechtenstein	175	0.81	46.73	3.15	3.66	513.85	522.41	451.99	0.39	15.31	0.81	1.11	91.63	87.40	94.75
Luxembourg	4483	0.79	43.84	3.34	3.51	442.23	481.94	487.94	0.41	17.55	1.41	1.41	84.11	101.81	93.19
Latvia	2719	0.51	48.83	4.73	4.61	451.53	490.40	444.13	0.50	18.19	0.73	0.87	100.27	98.91	98.92
Mexico	2567	0.56	43.22	2.78	3.03	394.19	540.09	454.30	0.50	17.10	1.33	1.40	78.28	89.78	88.49
Netherlands	1382	0.48	51.59	3.76	3.98	573.72	538.46	508.79	0.50	16.28	1.19	1.18	84.05	88.56	96.62
Norway	2307	0.98	53.95	4.20	4.14	498.75	488.76	507.14	0.13	15.32	1.10	1.11	87.08	101.65	93.61
New Zealand	2048	0.92	52.09	4.02	3.97	536.51	508.79	522.00	0.27	16.90	1.12	1.15	94.52	98.96	96.17
Poland	1976	1.00	44.72	4.25	4.02	460.09	487.21	474.39	0.00	15.09	0.97	1.01	96.19	94.25	91.66
Portugal	2545	0.55	44.59	3.14	3.19	458.85	489.84	463.23	0.50	16.09	1.26	1.29	85.61	94.61	88.79
Russian Federation	3719	0.73	49.72	4.80	4.69	478.71	493.53	464.20	0.44	17.05	0.62	0.75	98.14	97.59	90.91
Sweden	2464	0.97	50.38	4.40	4.29	509.90	504.44	501.87	0.17	16.16	0.99	1.05	88.61	83.09	89.80
United States	3010	0.59	52.50	4.65	4.66	492.56	484.51	483.13	0.49	16.27	0.85	0.87	96.10	99.17	91.62

Annex 2 – PISA 2003, descriptive statistics

		Mean							Standard deviation						
Country	Nobs	Share of pupils attending ref. grade	Parental SES index	Mother degree	Father degree	Score in Math	Score in reading	Score in Science	Share of pupils attending ref. grade	Parental SES index	Mother degree	Father degree	Score in Math	Score in reading	Score in Science
Australia	12551	0.91	52.61	3.49	3.54	522.33	523.85	522.78	0.28	16.04	1.38	1.37	94.01	93.16	98.19
Austria	4597	0.52	47.42	3.25	3.65	511.86	497.09	497.37	0.50	16.08	1.03	1.14	88.47	94.67	90.21
Belgium((Fl)	5059	0.73	50.81	3.59	3.65	552.56	528.99	528.02	0.45	16.76	1.36	1.33	101.97	94.44	93.88
Belgium (Fr)	3737	0.59	50.56	3.72	3.70	506.97	486.94	490.24	0.49	16.85	1.46	1.43	101.17	102.22	100.22
Brazil	4452	0.66	40.53	2.80	2.83	360.41	406.90	393.06	0.47	15.96	1.78	1.79	91.67	97.37	85.53
Canada	27953	0.80	50.75	3.83	3.65	521.40	516.18	508.65	0.40	15.97	1.17	1.25	84.03	84.49	92.66
Switzerland	8420	0.83	48.09	3.02	3.39	518.24	491.76	502.81	0.38	15.83	1.26	1.38	91.50	86.07	96.61
Czech Republic	6320	0.54	51.62	3.36	3.46	534.95	505.64	541.17	0.50	14.72	0.88	0.91	95.58	88.83	96.23
Germany	4660	0.82	49.60	3.08	3.41	508.41	497.12	508.23	0.39	16.26	1.32	1.43	97.16	100.74	103.44
Denmark	4218	0.91	49.13	3.76	3.51	513.69	491.32	474.54	0.29	15.48	1.34	1.29	87.10	81.02	93.98
Spain	10791	0.74	44.92	2.81	2.95	494.78	489.91	490.43	0.44	16.84	1.64	1.70	82.15	86.52	89.55
Finland	5796	0.87	50.76	3.92	3.70	542.81	541.60	544.49	0.33	16.93	1.33	1.43	79.50	73.65	83.32
France	4300	0.63	49.02	3.17	3.25	514.73	500.04	515.89	0.48	16.74	1.33	1.35	87.03	88.77	101.88
United Kingdom	9535	0.84	49.54	3.48	3.32	514.44	512.24	519.38	0.36	16.56	1.24	1.29	88.08	87.81	97.03
Greece	4627	0.90	46.38	3.02	3.10	440.88	468.10	477.49	0.30	16.86	1.38	1.50	89.83	95.80	90.87
Hong Kong-China	4478	0.60	41.16	1.95	2.16	555.86	513.87	544.61	0.49	13.44	1.24	1.28	94.18	76.87	85.75
Hungary	4765	0.92	48.33	3.33	3.35	488.59	480.66	501.54	0.28	15.26	1.05	0.96	89.66	84.97	89.84
Indonesia	10761	0.87	35.14	2.03	2.38	361.51	383.97	397.19	0.34	18.16	1.51	1.55	73.05	64.99	56.94
Ireland	3880	0.97	48.49	3.22	3.10	504.68	517.21	507.12	0.17	15.81	1.29	1.40	81.76	81.51	86.68
Iceland	3350	1.00	53.63	3.14	3.33	515.05	491.78	494.79	0.00	16.72	1.27	1.21	86.39	91.13	88.31
Italy	11639	0.84	47.54	3.04	3.04	496.00	500.99	515.11	0.37	16.29	1.27	1.27	89.79	90.84	96.56
Japan	4707	1.00	49.84	3.78	3.67	533.51	497.36	546.98	0.00	14.74	1.21	1.34	96.71	98.60	102.43
Korea	5444	0.99	46.09	2.94	3.33	540.60	532.85	536.84	0.12	13.45	1.33	1.41	89.23	76.91	93.98
Liechtenstein	332	0.79	50.80	2.90	3.51	536.46	525.66	525.81	0.41	14.97	1.21	1.36	95.28	83.80	96.65
Luxembourg	3923	0.85	48.18	3.22	3.42	493.48	479.78	483.07	0.36	16.60	1.71	1.62	88.01	93.51	96.07
Latvia	4627	0.81	50.74	4.09	3.95	486.17	493.02	491.39	0.39	16.52	1.04	1.07	83.35	82.24	84.33
Macao-China	1250	0.56	39.88	1.80	1.92	522.79	493.66	521.21	0.50	12.64	1.32	1.28	84.57	64.51	81.35
Mexico	29983	0.76	41.73	2.25	2.51	405.40	421.72	421.79	0.43	18.76	1.73	1.77	74.47	77.67	72.07
Netherlands	3992	0.51	51.48	3.17	3.46	542.12	516.89	528.71	0.50	15.87	1.35	1.43	89.80	80.60	93.86
Norway	4064	0.99	54.68	3.88	3.85	495.64	499.68	484.63	0.08	15.38	1.17	1.20	88.35	95.18	95.94
New Zealand	4511	0.93	51.62	3.44	3.26	525.62	523.40	523.03	0.25	16.41	1.38	1.37	95.03	98.82	97.74
Poland	4383	0.96	44.77	3.23	3.12	489.00	495.19	496.26	0.19	14.87	0.90	0.98	86.42	88.88	94.04
Portugal	4608	0.64	42.95	2.05	2.06	465.23	476.10	466.71	0.48	15.98	1.82	1.77	83.98	86.68	86.26
Russian Federation	5974	0.70	50.22	3.67	3.59	472.44	446.89	493.71	0.46	16.76	0.97	0.95	88.04	83.10	88.95
Slovak Republic	7346	0.62	49.60	3.29	3.38	504.12	475.22	501.58	0.48	16.21	0.89	0.93	88.91	84.69	92.89
Sweden	4624	0.97	50.71	3.83	3.57	507.95	513.12	505.00	0.16	16.16	1.35	1.44	90.97	89.01	98.30
Thailand	5236	0.57	37.19	1.86	1.99	422.73	426.33	435.52	0.50	16.20	1.39	1.39	80.80	72.73	75.58
Tunisia	4721	0.38	37.49	1.46	2.06	359.34	375.24	385.33	0.49	17.80	1.45	1.49	77.24	84.58	76.00
Turkey	4855	0.94	41.90	1.63	2.36	426.72	443.52	436.14	0.24	15.33	1.39	1.53	97.81	84.79	85.89
Uruguay	5835	0.58	46.10	3.03	2.98	412.99	422.68	429.23	0.49	18.17	1.72	1.72	99.08	116.26	101.95
United States	5456	0.68	54.19	3.62	3.53	481.47	494.09	490.01	0.47	16.38	1.19	1.22	90.38	94.29	93.96
Yougoslavia	4405	1.00	48.30	3.60	3.71	436.36	411.01	436.08	0.00	16.81	1.21	1.17	81.33	74.49	74.67

Annex 3 – PISA 2006, descriptive statistics

		Mean							Standard deviation						
Country	Nobs	Share of pupils attending ref. grade	Parental SES index	Mother degree	Father degree	Score in Math	Score in reading	Score in Science	Share of pupils attending ref. grade	Parental SES index	Mother degree	Father degree	Score in Math	Score in reading	Score in Science
Argentina	4339	0.71	45.98	2.88	2.72	388.12	383.93	398.33	0.45	16.95	1.87	1.84	90.14	110.10	92.48
Australia	14170	0.91	52.97	3.55	3.50	516.26	508.69	523.13	0.29	16.40	1.31	1.34	85.86	92.75	99.70
Austria	4927	0.52	48.13	3.39	3.77	509.51	493.95	513.86	0.50	16.60	1.09	1.18	91.52	101.68	93.29
Azerbaijan	5184	0.94	50.32	3.77	3.95	476.76	354.98	385.35	0.23	18.71	1.20	1.25	44.35	66.03	53.27
Belgium((Fl)	5124	0.75	49.80	3.73	3.67	545.82	524.34	531.35	0.44	16.26	1.32	1.32	93.18	98.87	88.91
Belgium (Fr)	3733	0.55	50.62	3.64	3.66	500.99	483.55	495.68	0.50	17.01	1.43	1.43	97.97	102.14	98.50
Bulgaria	4498	0.96	48.03	3.48	3.38	417.41	406.83	439.05	0.21	16.26	1.11	1.02	93.71	109.81	100.77
Brazil	9295	0.59	42.52	2.56	2.45	365.57	389.18	385.25	0.49	18.42	1.82	1.85	87.27	94.26	85.97
Canada	22646	0.84	52.60	3.99	3.76	517.42	512.32	522.52	0.37	15.75	1.18	1.26	82.21	95.24	92.37
Switzerland	12192	0.84	48.53	3.18	3.52	528.29	496.60	508.02	0.36	15.84	1.35	1.44	90.62	87.57	93.68
Chile	5233	0.78	41.27	2.82	2.91	417.08	447.86	443.11	0.41	16.76	1.49	1.50	82.41	94.79	87.38
Colombia	4478	0.61	43.04	2.61	2.69	373.83	390.31	391.86	0.49	17.36	1.90	1.94	82.77	96.91	80.85
Czech Republic	5932	0.50	50.76	3.52	3.55	536.03	509.64	537.61	0.50	14.70	0.99	0.99	103.75	108.19	99.67
Germany	4891	0.84	49.19	3.25	3.50	504.32	496.53	516.21	0.37	16.35	1.27	1.34	95.49	107.99	97.09
Denmark	4532	0.89	49.25	3.91	3.64	512.23	493.80	494.72	0.32	17.10	1.32	1.29	80.01	84.90	89.69
Spain	19604	0.69	46.32	2.94	2.99	501.65	479.52	504.51	0.46	17.34	1.52	1.58	83.89	82.58	84.08
Estonia	4865	0.73	50.81	3.82	3.66	516.77	502.38	533.73	0.45	16.56	1.05	1.08	75.96	81.39	80.32
Finland	4714	0.89	48.87	4.19	3.91	548.99	547.08	563.38	0.31	16.98	1.29	1.43	76.38	76.77	82.19
France	4716	0.60	48.78	3.24	3.25	496.43	488.66	496.12	0.49	16.60	1.32	1.37	91.38	98.95	98.24
United Kingdom	13152	0.97	50.06	3.69	3.52	497.27	495.64	514.27	0.18	16.23	1.19	1.25	84.01	96.01	102.80
Greece	4873	0.95	48.92	3.24	3.27	462.04	461.91	476.64	0.21	16.82	1.37	1.47	86.51	97.01	87.94
Hong Kong-China	4645	0.64	42.77	2.16	2.26	551.39	538.95	546.09	0.48	13.57	1.27	1.34	88.43	76.99	87.43
Croatia	5213	1.00	46.65	3.46	3.52	467.32	477.55	493.65	0.03	15.09	1.12	1.09	79.02	84.93	82.26
Hungary	4490	0.96	48.20	3.41	3.37	496.18	488.10	508.72	0.20	15.16	1.09	1.02	85.38	87.53	83.67
Indonesia	10647	0.89	37.96	1.98	2.28	380.69	383.92	384.76	0.32	15.66	1.46	1.51	69.47	64.73	59.36
Ireland	4585	0.97	49.22	3.40	3.29	502.34	518.65	509.49	0.16	16.37	1.27	1.36	77.67	87.34	90.86
Iceland	3789	1.00	54.01	3.33	3.51	505.59	484.99	490.95	0.04	16.99	1.38	1.31	83.64	92.19	93.50
Israel	4584	0.86	53.27	3.88	3.94	443.32	441.30	455.63	0.35	15.99	1.35	1.36	102.41	113.34	107.85
Italy	21773	0.82	46.38	2.91	2.89	473.63	477.01	487.15	0.38	16.32	1.14	1.17	92.29	102.63	93.27
Jordan	6509	0.92	51.71	3.02	3.32	389.18	409.49	427.10	0.28	17.19	1.64	1.63	75.92	85.85	83.72
Japan	5952	1.00	50.35	3.91	3.96	525.55	500.21	533.72	0.00	14.70	1.06	1.10	86.29	96.87	96.05
Kyrgyzstan	5904	0.93	47.09	4.22	4.27	315.90	290.54	326.33	0.26	18.02	1.15	1.12	80.05	94.27	77.78
Korea	5176	0.99	49.99	3.26	3.60	547.17	556.06	521.92	0.10	13.42	1.15	1.26	88.78	84.79	87.02
Liechtenstein	339	0.83	51.16	3.13	3.58	524.86	510.74	522.25	0.37	15.50	1.27	1.41	88.50	91.50	94.20
Lithuania	4744	0.88	49.59	3.99	3.70	485.61	469.33	486.52	0.33	17.87	1.05	1.06	85.35	91.38	87.00
Luxembourg	4567	0.88	47.69	2.94	3.15	490.49	480.07	486.85	0.33	16.62	1.77	1.69	88.86	95.43	93.60
Latvia	4719	0.82	49.34	3.87	3.64	491.24	484.86	493.78	0.38	16.63	1.07	1.06	77.00	84.48	80.31
Macao-China	4760	0.37	41.91	1.84	1.99	524.41	490.64	509.46	0.48	13.93	1.30	1.30	79.68	72.31	75.56
Mexico	30971	0.78	43.58	2.43	2.66	420.70	427.36	422.64	0.42	18.57	1.78	1.83	72.03	81.48	71.97
Montenegro	4455	1.00	48.89	3.83	3.93	395.84	388.23	408.79	0.04	16.14	1.21	1.16	79.63	84.98	75.79
Netherlands	4871	0.51	52.05	3.41	3.59	537.41	513.91	530.76	0.50	15.68	1.37	1.39	83.69	90.09	91.00
Norway	4692	0.99	53.11	4.00	3.90	489.84	484.37	486.93	0.07	15.34	1.17	1.21	86.71	99.50	91.99
New Zealand	4823	0.94	51.79	3.60	3.46	523.77	522.74	532.68	0.24	15.97	1.38	1.33	88.49	99.79	103.74
Poland	5547	0.97	45.32	3.25	3.18	500.95	512.63	503.29	0.18	15.33	0.84	0.79	84.35	95.36	87.82
Portugal	5109	0.54	42.01	2.01	1.93	470.94	476.84	478.97	0.50	16.30	1.78	1.76	85.43	93.14	84.17
Qatar	6265	0.77	61.67	3.24	3.60	317.74	312.51	349.08	0.42	12.97	1.79	1.65	83.61	101.25	78.49
Romania	5118	0.95	43.39	3.65	3.65	414.97	391.97	416.61	0.22	16.25	1.22	1.18	79.79	88.77	77.84
Russian Federation	5799	0.66	51.33	3.65	3.54	478.66	442.37	481.50	0.47	17.09	0.97	0.96	84.46	85.79	85.86
Serbia	4798	0.99	48.51	3.56	3.64	436.64	402.86	436.93	0.10	16.28	1.18	1.15	86.28	86.04	80.93
Slovak Republic	4731	0.61	47.50	3.34	3.36	495.10	470.55	491.22	0.49	15.80	0.89	0.91	89.62	98.83	89.57
Slovenia	6595	0.99	47.83	3.29	3.23	482.21	468.58	494.19	0.08	15.71	1.07	1.00	84.70	88.96	93.79
Sweden	4443	0.98	50.70	4.03	3.74	503.23	508.99	504.23	0.14	15.86	1.29	1.40	85.46	93.35	91.31
Chinese Taipei	8815	0.69	49.55	2.80	2.97	562.75	506.68	543.71	0.46	15.91	1.00	1.11	94.56	76.92	88.44
Thailand	6192	0.67	38.90	1.90	2.07	425.47	425.19	429.73	0.47	16.65	1.47	1.50	81.29	78.79	79.22
Tunisia	4640	0.49	37.85	1.84	2.43	363.91	378.96	384.19	0.50	18.74	1.60	1.63	85.47	87.79	77.22
Turkey	4942	0.57	39.83	1.55	2.22	428.25	452.92	427.61	0.50	15.47	1.27	1.45	89.33	83.21	79.55
Uruguay	4839	0.70	45.82	3.14	3.02	435.47	424.68	437.68	0.46	18.66	1.81	1.84	93.53	111.99	91.32
United States	5611	0.88	52.46	3.78	3.67	474.72		488.29	0.32	16.78	1.28	1.31	85.46		102.37

Annex 4 – Synthetic control as an indentifying strategy

Suppose we observe *i=1 to J* educational systems during *T* periods. Suppose that the first one (*i.e.* the French-Speaking Community of Belgium) is exposed to the intervention/policy change of interest in time T_0.

Let $Y_{i,t}$ be the outcome that could be observed for system *i* at time *t*

$$Y_{i,t}= \delta_t + \alpha_{it} D_{it} + \upsilon_{it} \qquad [1]$$

$$\upsilon_{it} \equiv Z_i\theta_t + \lambda_t\mu_i + \varepsilon_{it}$$

with $D_{it}=1$ if $i=1$ and $t>T_0$ and $D_{it}=0$ otherwise, where δ_t is a common time period effect, Z_i is a vector of observed covariates that potentially influence the outcome, μ_i is an unobserved system-specific effect, λ_t is an unknown common factor, and ε_{it} are unobserved transitory shocks at the system level with zero mean for all.

We aim at estimating for $t>T_0$ (*i.e.* after the intervention)

$$\alpha_{1t}= Y_{1t} – Y^N_{1t}$$

Because Y_{1t} is observed, we only need to estimate its counterfactual Y^N_{1t}.

Consider a (*Jx1*) vector of weights $W=(w_2,....w_J)$ such that $w_i \geq 0$ and $w_2+....+w_J=1$. Each particular vector *W* represents a potential synthetic control (SC), that is, a particular weighted average of control systems.

Consider an arbitrary linear combination *K* of all pre-intervention outcomes $Y^K_i \equiv \sum^{T0}_{s=1} (k_iY_{is})$

$$\sum^J_2(w_i Y^K_i) - Y^K_1= \sum^J_2 w_i \sum^{T0}_1 (k_iY_{is}) - \sum^{T0}_1 (k_iY_{1s}) \qquad [2]$$

Using [1], this can be written

$$\sum^J_2(w_i Y^K_i) - Y^K_1= \sum^{T0}_1 k_s \delta_s + \sum^J_2 w_i\sum^{T0}_1 k_i \upsilon_{is} - \sum^{T0}_1 k_s \delta_s - \sum^{T0}_1 k_s \upsilon_{1s} \qquad [3]$$

or equivalently, exploiting the fact that $\sum^J_2 w_i =1$

$$\sum^J_2(w_i Y^K_i) - Y^K_1= \sum^J_2 w_i (\upsilon_{is} - \upsilon_{1s}) \qquad [4]$$

Using the definition of υ_{is} and υ_{1s} in [1], the expression becomes

$$\sum^J_2(w_i Y^K_i) - Y^K_1=(\sum^J_2 w_i Z_i – Z_1)(\sum^{T0}_1 k_s\theta_s) + (\sum^J_2 w_i \mu_i - \mu_1)(\sum^{T0}_1 k_s\lambda_s) + \sum^J_2 w_i\sum^{T0}_1 k_s(\varepsilon_{is} - \varepsilon_{1s})$$

[5]

In addition (for any *t*),

$\sum^{J}_{2}(w_i Y_{it}) - Y^{N}_{1t}=(\sum^{J}_{2}w_i Z_i - Z_1)\theta_t + (\sum^{J}_{2}w_i \mu_i - \mu_1)\lambda_t + \sum^{J}_{2}w_i(\varepsilon_{it} - \varepsilon_{1t})$ [6]

Suppose that we choose $W^{*}=(w^*_{2,}....w^*_{J})$ such that

$$\sum^{J}_{2}(w^*_i Y^{K}_{i}) - Y^{K}_{1}=0 \text{ and } \sum^{J}_{2}w^*_i Z_i - Z_1=0$$

Then, the left-hand term of [5] as well as the first term of the right-hand part of [5] and [6] disappear. What is more, if $\sum^{T_0}_{1}k_s\lambda_s \neq 0$, we obtain from [5] that

$\sum^{J}_{2}w_i \mu_i - \mu_1 = -[1/(\sum^{T_0}_{1}k_s\lambda_s)]\sum^{J}_{2}w_i\sum^{T_0}_{1}k_s(\varepsilon_{is} - \varepsilon_{1s})$ [7]

Hence [6] becomes a function of the random error terms ε exclusively, with an expected value equal to zero

$\sum^{J}_{2}(w^*_i Y_{it}) - Y^{N}_{1t=} - \lambda_t [/(\sum^{T_0}_{1}k_s\lambda_s)]\sum^{J}_{2}w_i\sum^{T_0}_{1}k_s(\varepsilon_{is} - \varepsilon_{1s}) + \sum^{J}_{2}w_i(\varepsilon_{it} - \varepsilon_{1t})$ [8]

Therefore, for $t>T_0$ we have that $\sum^{J}_{2}(w^*_i Y_{it})$ equates the (unobserved) counterfactual Y^{N}_{1t}. Hence,

$\alpha_{1t}= Y_{1t} - \sum^{J}_{2}(w^*_i Y_{it})$

The computation of W^* is done by minimizing the pseudo-distance $|| X_1 - X_{SC}W||$ subject to the condition that $w_i \geq 0$ and $w_2+....+w_J=1$ where $X_1= (Z_1, Y^{K1}_{1,...,} Y^{KM}_{1,})$ is the vector of pre-treatment characteristics that comprises Z_1= observable controls and $Y^{K1}_{1,...,} Y^{KM}_{1}$ *i.e.* linear combinations of the PISA scores. The similar vector for the non-treated countries is X_{SC}.

Annex 5 – Inference analysis with Synthetic control

Unlike Abadie, Diamond and Hainmueller (2007) we have access to individual data within each country. Like them, we run the Stata *synth* procedure, using data aggregated at the country level (*Y, Z*). This explains that we rely on (numerous) individual data to do hypothesis testing and computed the results reported in Tables 3 and 4.

The statistics we aim at are standard *t-tests* gauging the plausibility that two means (the post-treatment for the treated country and its synthetic control) are statistically different.

$t= (Y^1\text{-}Y^{SC}) / (Var^2(1/N_1 + 1/N_{SC}))^{1/2}$ [1]

with Var^2 the pooled sample standard deviation equal to

$Var^2= Var^1\ (N_1\text{-}1) + Var^{SC}(N_{SC}\text{-}1)/(N_1+ N_{SC}\text{-}2)$

where S^1 is the standard deviation characterising the treated entity post treatment (here the French Speaking Community of Belgium in 2006) and Var^{SC} the standard deviation characterizing its synthetic equivalent. It is important to stress how the latter is computed.

Assume we have $N_1,N_2\dots\ N_J$ students *j* in each of the *J* countries that participated to PISA, with N_1 designating the sample size for the treated country (i.e. the French-Speaking Community of Belgium).

The synthetic control score for the post-treatment period computed by the STATA code developed by Abadie, Diamond and Hainmueller (2007) uses country-level averages $Y_i =1/N_i \sum^{Ni}_{1=1} Y_{il.}$ The delivered score is equal to

$Y^{SC}=\sum^J_{i=2} w*_i\, Y_i = \sum^J_{i=2} w*_i(1/N_i \sum^{Ni}_{1=1} Y_{il})$ [2]

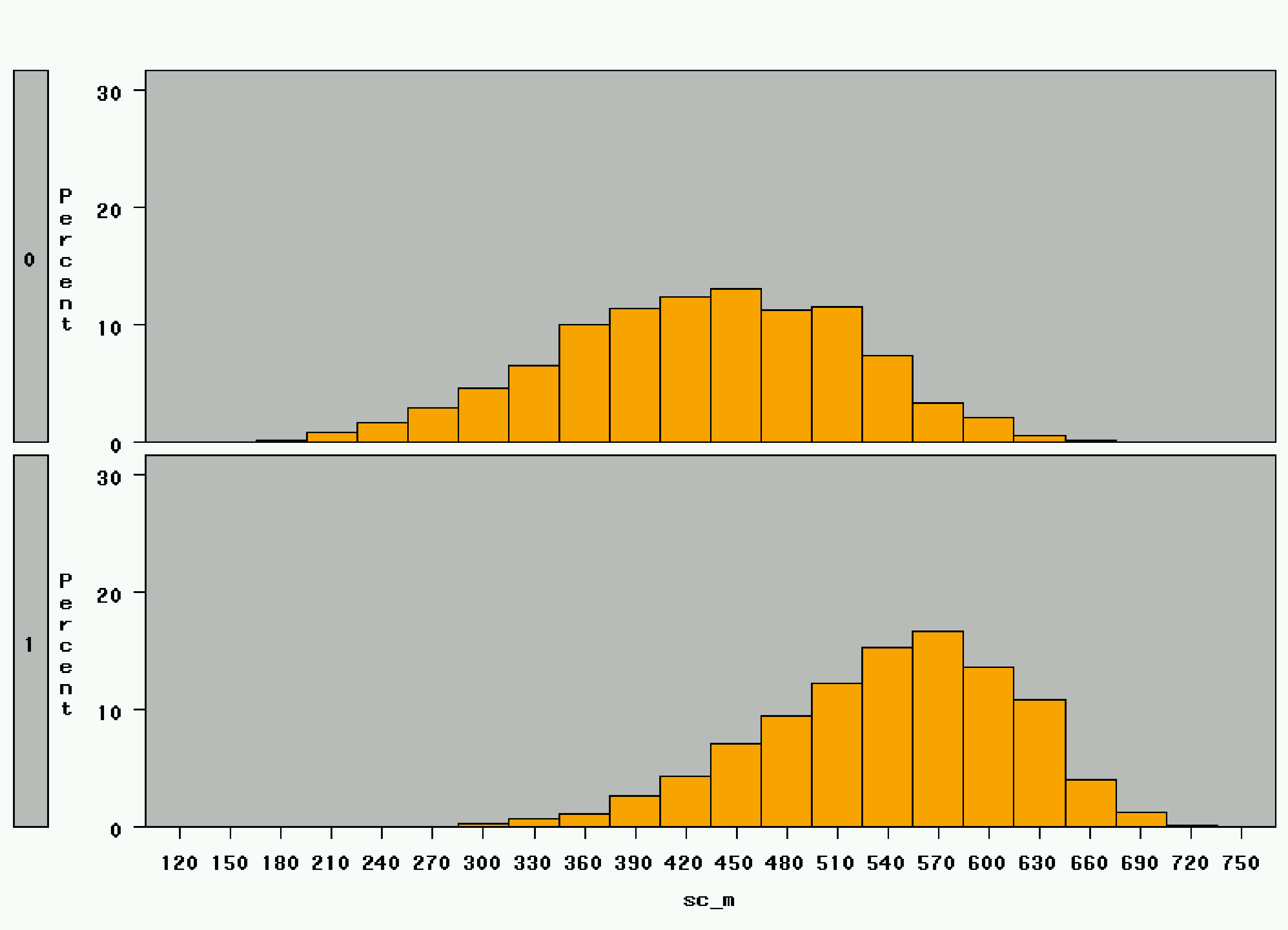
The point is that the latter can be replicated using individual/disaggregated data by *i)* applying the estimated weights *W** to the entire sample $N^{\$}=N_2+\dots + N_J$ of individuals forming the synthetic control entity ($w*_i\ Y_{ij}$) *ii)* provided the individual values are weighted by $N^{\$}/N_i$

$Y^{SCR}= 1/N^{\$}(\sum^J_2 \sum^{Ni}_1\ \theta*_i\ Y_{il})$ with $\theta*_i \equiv N^{\$}/N_i\ \ w*_i$ [3]

It is immediate to show that [3] is equal to [2]. What is more, the same weighing strategy can be used to compute from individual data the variance characterising the synthetic control entity.

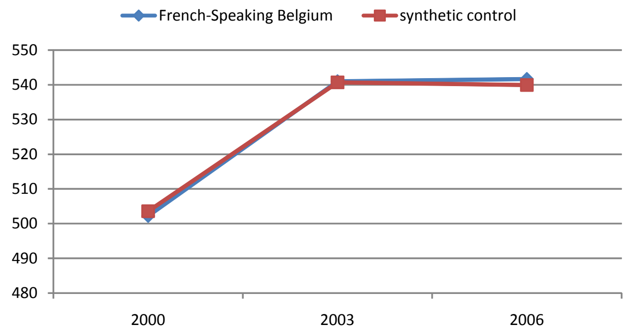
$S^{SC} = 1/N^{\$}(\sum^J_2 \sum^{Ni}_1\ (\theta*_i\ Y_{ij} - Y^{SCR})^2$ [4]-

Annex 6 – Pisa score distribution by grade (on the vertical axis, 0= below grade 10 and 1= grade 10,)

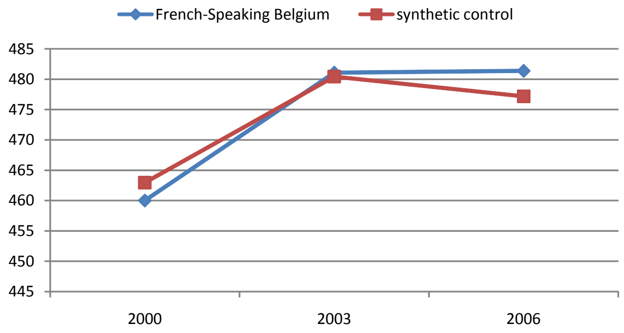


Annex 7 – PISA average score[25] in 2000, 2003 (before treatment) vs 2006 (after treatment). Comparison between the French-Speaking Community Belgium and its synthetic control.

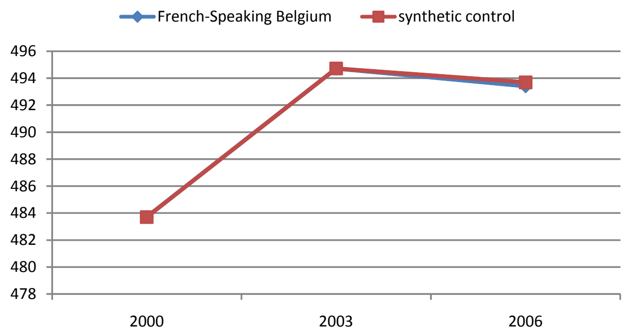
Grade 10 only



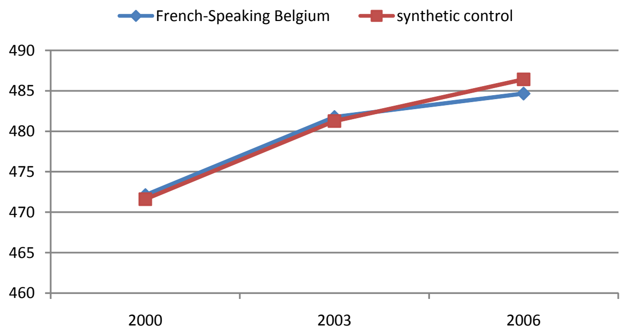
Grade 10 (<70 th perc)



All grades pooled



All grades pooled (>70 th perc <=30 th perc)



[25] Country/entity averages, based on individual unweighted average score in math, science and reading PISA scores.

Annex 8 – Control/predictor variables [26]. Comparison between the French-Speaking Community Belgium and its synthetic control (all grades pooled).

Year	Synthetic control (all grades pooled)							French-Speaking Community of Belgium						
	Share of pupils attending the reference grade	Student teacher ratio	Ratio of computers to school size	Share of teachers with proper certification	Highest parental socio-economic index (HISEI)	Mother education	Father education	Share of pupils attending the reference grade	Student teacher ratio	Ratio of computers to school size	Share of teachers with proper certification	Highest parental socio-economic index (HISEI)	Mother education	Father education
2000	0.63	13.90	0.09	0.81	48.92	4.41	4.31	0.59	10.06	0.07	0.77	50.67	4.05	4.20
2003	0.67	12.61	0.10	0.90	50.40	3.51	3.50	0.59	10.14	0.09	0.86	50.56	3.72	3.70
2006	0.64	13.12	0.09	0.91	50.27	3.52	3.45	0.55	9.90	0.11	0.78	50.62	3.64	3.66

[26] Country/entity averages, based on individual unweighted average score in math, science and reading PISA scores.

Annex 9 – Country weights forming the synthetic French-Speaking Community of Belgium.

					Analysis of Grade 10 scores					
Country	all data	<=p10	<=p20	<=p30	<=p40	<=p50	<=p60	<=p70	<=p80	<=p90
AUS										
AUT									0.140	
BFL	0.253					0.455	0.330	0.439	0.116	0.138
BRA										
CAN										
CHE										
CZE	0.526							0.244		0.763
DEU										
DNK										
ESP										
FIN										
FRA			0.346		0.491		0.240		0.260	
GBR										
GRC										
HUN										
IRL										
ISL						0.100				
ITA										
JPN										
KOR										
LIE				0.850		0.210				
LUX										
LVA	0.221	0.124	0.444	0.570	0.451	0.331	0.430	0.318	0.479	0.100
MEX										
NLD		0.226	0.140	0.480	0.580	0.400			0.366	
NOR										
NZL										
POL										
PRT		0.425	0.700							
RUS		0.226								
SWE										
USA										
Sum	1	1	1	1	1	1	1	1	1	1

	Analysis of overall (all grades pooled) scores									
	all	>p10	>p20	>p30	>p40	>p50	>p60	>p70	>p80	>p90
Country	data	<=p90	<=p80	<=p70	<=p60	<=p50	<=p40	<=p30	<=p20	<=p10
AUS										
AUT										
BFL										
BRA										
CAN										
CHE										
CZE	0.330	0.363	0.386	0.481	0.533	0.532	0.534	0.510	0.438	0.345
DEU										
DNK										
ESP										
FIN										
FRA										
GBR										
GRC										
HUN										
IRL										
ISL	0.270	0.300	0.250	0.290						
ITA										
JPN										
KOR										
LIE										
LUX					0.280	0.280				
LVA	0.169	0.231	0.323	0.235	0.580	0.184	0.990	0.500	0.500	
MEX					0.100					
NLD	0.210	0.500	0.460	0.300				0.220	0.530	
NOR										
NZL										
POL										
PRT	0.670	0.530	0.600	0.610	0.170	0.650	0.590	0.670	0.190	0.584
RUS	0.386	0.299	0.160	0.192	0.364	0.191	0.380	0.359	0.349	0.710
SWE										
USA										
Sum	1	1	1	1	1	1	1	1	1	1