Economics of Education Review 41 (2014) 137–160

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The distributional impacts of a universal school reform on mathematical achievements: A natural experiment from Canada

Catherine Haeck, Pierre Lefebvre *, Philip Merrigan

Department of Economics, Université du Québec à Montréal, CP 8888, Succ. Centre-ville, Montréal, QC, Canada H3C 3P8

ARTICLE INFO

Article history:
Received 26 July 2013
Received in revised form 11 March 2014
Accepted 17 March 2014
Available online 2 April 2014

JEL classification:
A21
I22

Keywords:
Natural experiment
Teaching reform
Math scores
Distributional effect

ABSTRACT

We investigate the impact of an ambitious provincial school reform in Canada on students' mathematical achievements. It is the first paper to exploit a universal school reform of this magnitude to identify the causal effect of a widely supported teaching approach on students' math scores. Our data set allows us to differentiate impacts according to the number of years of treatment and the timing of treatment. Using the changes-in-changes model, we find that the reform had negative effects on students' scores at all points on the skills distribution and that the effects were larger the longer the exposure to the reform.
© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Empirical research has shown that measures of educational attainment alone may not be sufficient to capture the extent to which human capital triggers economic growth and impacts individual labor market outcomes. Research shows that concrete measures of academic achievement and cognitive skills, along with educational attainment, are strongly correlated with labor market outcomes, such as earnings and unemployment (Currie & Thomas, 2001; Hanushek & Woessmann, 2008; Murnane, Willett, Duhaldeborde, & Tyler, 2000; Murnane, Willett, & Levy, 1995; Neal & Johnson, 1996).[1] A number of studies have documented the specific importance of mathematical abilities in adult socioeconomic success (e.g. Ingram & Neumann, 2006; Murnane et al., 1995; Rose & Betts, 2004). Evidence suggests that skills in mathematics have long lasting effects.[2]

Developing these skills should be of great policy interest, and so should understanding which policy may

* Corresponding author. Tel.: +1 514 987 3000x8373; fax: +1 514 987 8494.
E-mail addresses: Haeck.catherine@uqam.ca (C. Haeck), Lefebvre.pierre@uqam.ca (P. Lefebvre), Merrigan.philip@uqam.ca (P. Merrigan).

http://dx.doi.org/10.1016/j.econedurev.2014.03.004
0272-7757/© 2014 Elsevier Ltd. All rights reserved.

[1] Other recent studies show that non-cognitive skills (i.e. behavioral and social skills) also play an important role in predicting labor market outcomes. Although non-cognitive skills are more difficult to measure, they seem more malleable over the life cycle (Heckman & Rubinstein, 2001; Heckman, Stixrud, & Urzua, 2006).

[2] Murnane et al. (1995) find that math test scores measuring basic knowledge of mathematics (such as the working of fractions and decimals) predict future wages better than reading and vocabulary test scores. They also show that the effect on wages is larger 6 years after graduation than 2 years after.

help (or hinder) the development of these skills. A large body of research in the literature on human capital has investigated the impact of different inputs in the educational production function on achievement. Teacher quality has been shown to be of great importance in predicting the success of students.[3] Other types of resources (e.g. per student expenditure, school facilities, class size) have generally been shown to be poor predictors of student performance.[4] While the impact of school resources has been widely studied, few studies in economics have addressed the importance of what is being taught and how it is being taught.[5]

In this paper, we estimate the impact of Québec's (the second most populated province in Canada) ambitious and universal school reform implemented in the early 2000's on children's mathematical ability throughout primary and secondary school. At the time of the reform, the performance of students in the province of Québec was comparable to that of students from the top performing countries in international assessments. Nonetheless, the educational system in Québec was still subject to severe criticism at home due to its alarmingly large high school dropout rate, especially among male students.[6] To ensure the success of all students, the province decided to implement an ambitious reform introducing a new program in each and every school across the province which drastically changed the way teaching was delivered to all children in primary and secondary schools. The Québec education program (MELS, 2001, 2003, 2007) relied on a socio-constructivist teaching approach focused on problem-based and self-directed learning. This approach mainly moved teaching away from the traditional/academic approaches of memorization, repetitions and activity books, to a much more comprehensive approach focused on learning in a contextual setting in which children are expected to find answers for themselves.

Evidence from Bulle (2011) suggests that most OECD countries are moving away (or have long moved away) from the traditional (more academic) teaching approach. More specifically, the teaching approach promoted by the Québec reform is comparable to the reform-oriented teaching approach in the United States. As of 2006, this approach was widely spread across the United States (although more traditional approaches remained dominant) and it was supported by leading organizations such as the National Council of Teachers of Mathematics, the National Research Council, and the American Association for the Advancement of Science. Yet few studies in economics have addressed the impact of various teaching approaches, let alone the approach promoted by the Québec reform.

The Québec reform/experiment provides important advantages for the purpose of evaluation. First, Québec's Department of Education implemented the reform and all schools (public and private) were compelled to apply the new education program. Combined with a rich data set, this allows us to estimate the impact of a universal reform, on students of different abilities. Second, because teaching in the Rest of Canada (RofC) continued to be delivered in the same way throughout the period, it is possible to estimate the effect of the reform using students from the other provinces as controls. Third, the reform combined with our data set allows us to observe students treated for up to nine years (as opposed to three and less in the studies cited below). This longer observation period also allows us to estimate (to some extent) the long term effects of the reform.[7] Fourth, although the treatment period is long, because the reform was implemented in steps, all of our estimates are derived from a comparison of students in the same grade close by in time, which reduces the possible bias induced by other reforms taking place at the same time.

We use the Canadian National Longitudinal Survey of Children and Youth (NLSCY) for the analysis, which provides students' test scores in mathematics. We estimate the effect of the reform on a standardized measure of mathematical ability using two econometric methods. First, we apply the 'Difference-In-Differences' (DID) framework, a method largely used for evaluating the effects of policy changes (Angrist & Krueger, 1999; Blundell & Costa Dias, 2009). Second, the estimations are conducted with the 'Changes-in-Changes' (CIC) non parametric estimator, developed by Athey and Imbens (2006), which generalizes the DID model. Using the CIC framework, we assess the impact of the reform at different points on the skills distribution, from the least performing students to high achievers.

Studying this reform contributes to the literature by further identifying the determinants of mathematical abilities, and more specifically by identifying the causal impact of an increasingly popular teaching approach. It is the first paper to exploit a universal school reform of this magnitude to identify the causal effect of the teaching approach on the development of the mathematical skills of students. Our results suggest that the reform had negative effects on the development of students' mathematical abilities as measured by the CAT/2 test and that the effects were larger the longer the exposure to the reform and hence to the teaching method.

The outline of the paper is as follows. Section 2 highlights the distinctive features of the education system

[3] For example, among many others, Hanushek and Rivkin (2010) show that teacher quality (as well as principals and administrators) can be linked with children's test scores, such that better teachers can have significant effects on student achievement. Chetty, Friedman, and Rockoff (2011) show that the quality of teachers has large positive impacts not only on test scores, but also on long-term outcomes (e.g. attending college, future earnings, not having children as teenagers).

[4] In economics, see, for example, Angrist and Lavy (1999), Hanushek (2003), Hoxby (2000), and Rivkin, Hanushek, and Kain (2005). In educational research, see Teddlie and Reynolds (2000).

[5] One exception is Machin and McNally (2008) on the impact of a highly-structured hour of literacy on students reading and English skills. Other studies related to the teaching approach include, Angrist and Lavy (2002) on the use of Computer Aided Instruction, and Rouse and Krueger (2004) on the impact of instructional computer programs on literacy.

[6] In 1999, the dropout rate was 16.0% in Québec versus 12.0% in Canada, and 19.9% versus 14.7% for males at age 20 (Bowlby & McMullan, 2002).

[7] Borman, Hewes, Overman, and Brown (2003) found that greater impacts were estimated when the school reform evaluated had been in place for a greater number of years (more than 5 years).

in Québec compared to the other Canadian provinces, and describes the school reform implemented as of 2000. International research on comparable teaching approaches are also discussed. Section 3 exposes the econometric methodology used to identify the causal effect (treatment on the treated effect) of the school reform on achievement in math. Section 4 describes the data set used and presents descriptive statistics of key variables. The estimated effects of the reform are presented in Section 5. In Section 6, we provide further evidence on the impact of the reform using international assessment data and behavioral outcomes available in our main data set. The last section offers concluding remarks.

2. Québec's school system and teaching reform

In Canada, education is regulated and administered at the provincial level. The overall structure of the education system is comparable across all ten provinces, except for Québec where it is slightly different. In all of Canada, children start school in kindergarten at age 5, and then move on to primary school, where they complete six years of education (grades 1–6); children then pursue their education in high school. In Québec, high school consists of five years of education, grades 7–11, while in the RofC children must complete six years of education, grades 7–12, to obtain their high school diploma. Grades 7–11 in Québec and the RofC are comparable.[8]

Prior to and after the reform, teachers across Canada have the same level of qualifications when entering the profession: they must hold a Bachelor of Education degree from the faculty of education of a Canadian university, and must be certified to teach in the province. [9] More specifically, in Québec, the four year bachelor's degree in education has been a necessary condition to access the teaching profession since 1994.[10] Findings of this paper relate to the reform itself and the teaching method, and not to the teacher's level of training, or even more generally to teacher quality.[11]

We have limited information on teachers in the NLSCY and only until 2002 (many do not report their years of experience, sex, age or diplomas). The response rates are generally under 50%. We know that future teachers in their years of studies in education at the university level self-select according to their teaching preferences (kindergarten and primary or secondary). And through their university courses, they also choose their main teaching domain (e.g., arts, language, math, sciences). When they enter the profession in a specific school (first year of teaching with an interim certificate), it is clear that they are hired by School Boards for their potential teaching skills for a level and domain. Afterward, assignments are based on strict rules (years of experience) established in collective bargaining. We could not find any evidence that teachers with little experience were assigned to reform groups and there is no reason to think that such a strategy was adopted across the province.

We now provide a detailed overview of the reform and its implementation. Then we discuss quantitative research studying comparable reforms in other countries.

2.1. The reform

As of 2000, a comprehensive school reform impacting both primary and secondary schools was deployed all across the province of Québec. The reform aimed at making schools more responsive to the changing needs of children in order to improve their chances of success. Cross-curricular competencies and broad areas of learning[12] were introduced into the new program and formed the key elements of this new approach centering the teaching and learning environment around the students. More specifically, this approach was designed to enable students to "find answers to questions arising out of everyday experience, to develop a personal and social value system, and to adopt responsible and increasingly autonomous behaviors" (MELS, 2005).

In the classroom, students were expected to be more actively involved in their own learning and take responsibility for it. Critical to this aspect was the need to relate their learning activities to their prior knowledge and transfer their newly acquired knowledge to new situations in their daily lives. "Instead of passively listening to teachers, students will take in active, hands-on learning. They will spend more time working on projects, doing research and solving problems based on their areas of interest and their concerns. They will more often take part in workshops or team learning to develop a broad range of competencies." (MELS, 1999). This centralized approach in providing the program and training with a school-based execution is in many ways comparable to the current approach taken within the comprehensive school reform (CSR) models at the national level in the United States (Borman et al., 2003). The main differences are that in Québec, implementation was mandatory in each and every school, funding was not tied to the implementation, and training packages and support are centralized in many ways. These differences are critical: they imply that the reform had to be implemented in all schools, and that the resources and training was not tied to individual school characteristics. Whether private or public, English speaking or French speaking, all schools across the province were mandated to follow the reform according to the implementation schedule. This implies that all children in Québec were treated according to same timeline, and that parents were not able to self-select their children into or

[8] Our data set, further detailed in Section 6, covers grade 2–10 students.

[9] Holding a Bachelor's degree in Education gives an interim Certificate by the Minister of Education in each province. A permanent professional certificate is issued after completing the equivalent of one or two years of teaching in a setting where the courses, education programs and instructional materials used are prescribed, authorized or approved by the Minister of Education, and receiving a written recommendation from his or her superintendent or a school principal.

[10] The bachelor's degree in education is among the few programs that requires more than 3 years of study at the University level.

[11] From 1998 to 2010, the overall ratios of students to teachers decreased from 16.5 to 14.3 and the number of teachers did not change.

[12] A complete list of the competencies and areas of learning is provided in Table A.1 of the Appendix.

out of the reform, except by moving out of the province which they did not.

The school reform was planned at the highest level by civil servants at the Department of Education (MELS). The MELS imposes the program to be followed in each grade by every school. The 69 School Boards (60 Francophone and 9 Anglophone) responsible for all public schools, their superintendents and the school principals, are the channels and drive belts between the MELS and school teachers and students.

There was no formal mechanism or monitoring to insure enforcement. However, the very detailed education programs imposed by the reform pressured teachers to adapt their teaching methods. They were supported by education counsellors (at least one per school) who were responsible for helping teachers with program content, and teaching methods and approach. Furthermore, province-wide ministerial examinations were changed to reflect the new approach. For example, in grades 4 and 6 examinations are for "language reading and composition", and math in grade 6. At the secondary level, the exams assess progress in math, sciences, and history as well as citizenship education (grades 10 and 11) and French and English (teaching language and second language).[13] Evidently, the exams do not control the way teachers ensure progression of learning or skills. But the education programs expose in detail the targeted skills and broad areas of learning (see Table A.1). Given that the exams are tailored to the teaching approach, there is a strong incentive for teachers to follow the approach imposed by the MELS or else be faced with large rates of failure for these exams.

Extensive training was provided to support the new program. The year prior to the implementation in Elementary Cycle 1, teachers, principals and government officials began the task of preparing the implementation of the reform. Sixteen pilot schools along with several other Lead schools in the English sector experimented with the key concepts of the program of study, as well as school organizational approaches that could be best suited to the strategies required to maximize the effectiveness of the learning environment.

In June 2000, principals in conjunction with teachers began developing their implementation plans for September 2000. Each school was allowed to develop its own approach to deal with the implementation since no single approach was believed to meet the needs of each school across Québec. Teaching was organized by cycle. Some schools chose to organize teacher teams by cycle. Others opted for a "looping" model in which each teacher was assigned to one group of students for the entire cycle (e.g. grades 1 and 2). Some schools spent a lot of effort in developing themes and projects that actively involved the students, while others piloted a new reporting method to evaluate students that would be in tune with the new program. In 2000, all schools, both elementary and secondary, participated in some way to the development of the implementation of the reform despite the fact that it did not affect all levels of schooling at the time. Guides for teachers were produced. The implementation was staggered over many years (grades), giving time for teachers to adapt to the new programs. Unfortunately, the NLSCY does not provide any information on the extent to which the reform was implemented in the school attended by the child, but there is no evidence that the reform was not implemented in each and every school across the province.

Fig. 1 shows the implementation schedule of the reform. Students in grades 1 and 2 (Elementary Cycle 1) were introduced to the reform in September 2000. The changes were phased in for other cycles over time: September 2001 – grades 3 and 4 (Elementary Cycle 2); September 2003 – grades 5 and 6 (Elementary Cycle 3); September 2005 and 2006 – grades 7 and 8 respectively (Secondary Cycle 1); September 2007, 2008 and 2009 – grades 9, 10 and 11 respectively (Secondary Cycle 2). The original plan for grades 5–6 and secondary school was delayed by one year. While training for grade 5 and 6 teachers began as early as 2001, the implementation was delayed, from 2002 to 2003.

We are not aware of any econometric evaluation of the reform. Research by the Evaluation of the Reform at the Secondary School Level group (ERES, 2011a) on students starting high school in 2006–2007 revealed that students post-reform (relative to pre-reform students) perceived less favourably the classroom climate and some of the teaching practices. However, no credible control group was available for this study.

In sum, active competencies such as problem solving, strong communication skills, use of creativity, cooperation with others and teaching strategies based on the active participation of students were central to the reform, while more passive learning approaches such as memorization, repetition and traditional lectures in which teachers provide the content to be learned appears to have been downplayed by this teaching approach. Furthermore, because of the centralized nature of primary and secondary level education policies in Québec and teacher incentives, we are rather confident that the reform was applied uniformly in every school of the Province.

2.2. International comparability

Although the Québec reform is not unique (see for example the 1989 NCTM Standards in California, and the 1997 reform in Belgium's French Community) and the importance of mathematical skills is well documented, we are not aware of any study in the economics literature estimating the impact of the teaching approach promoted by these reforms on the development of students mathematical abilities. Bailey and Borooah (2010) provide a glimpse into what the effects may be. Using a multivariate approach, they estimate the influence of different factors (e.g. family type, education of the parent, minutes of instruction, etc.) on the mathematical skills of students. They use the OECD's Program for International Student Assessment (PISA) 2003 data on students from 41 countries. Their results suggest that relative to elaboration learning (the strategy promoted by the Québec reform),

[13] The MELS does not give access to administrative data sets on exam results. They make public each year the percentage of students who receive a passing grade by school board.

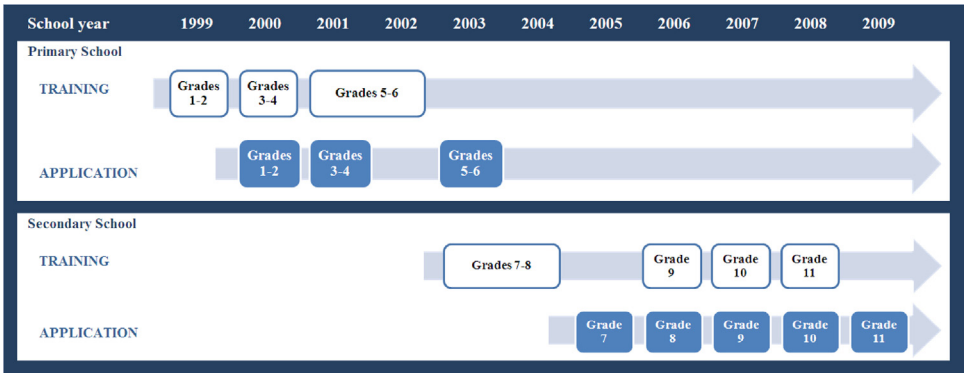


Fig. 1. Reform schedule and implementation.

memorization/rehearsal and control learning are positively correlated with the mathematics test score, but the estimated effects are small.

Studies in education research also provide some insights. In the United States, Comprehensive School Reforms (CSR) have been implemented by schools and districts for almost two decades to improve the country's many low-performing public schools. Although the CSR model developers are many and their designs differ, they typically employ similar strategies to achieve better student performance (American Institutes for Research – AIR, 2005): organizing the school to facilitate transformed teaching and learning, transforming curriculum and instruction, providing students with the necessary academic and social support, increasing teacher and principal effectiveness, as well as parental involvement. A meta-analysis of over 230 evaluations on 29 leading CSR models across the country found that their overall effects are positive, but small[14] (Borman et al., 2003).

Of the few experimental studies cited by Borman et al. (2003), one clearly relates to the Québec reform curriculum. Crawford and Snider (2000) study two curricula. The first one explicitly teaches mathematical concepts and focuses on the mastery of mathematical concepts through drill and repetition. The second one uses a more implicit approach in which the teacher sets up a situation in which the students have to learn and discover concepts through reasoning and discussions, but provides no explicit opportunities to review or practice. The latter shares similarities to the Québec reform. The authors find that the former approach (i.e. the more explicit approach) is more successful in producing mathematical knowledge.

In parallel with CSR, other reforms closely related to the Québec reform were conducted in the United States. Le et al. (2006) evaluate the impact of a more targeted approach, reform-oriented teaching, on students' achievements in mathematics and sciences in three school districts in the United States. Reform-oriented teaching promotes the active participation of students in their own learning. In this approach, inquiry-based activities are central: students are expected to ask questions, discuss alternative solutions, make connections between knowledge acquired in different subject areas and present the reasoning that led them to a preferred solution. Using multivariate analysis, the authors find that the relationship between reform-oriented teaching and achievement in mathematics and sciences is either non-significant or at best weakly positively significant.

Although reform-oriented teaching is gaining ground in the United States, there is limited empirical evidence concerning the effectiveness of different teaching approaches in raising students' achievements. The few studies listed above suggest that a more structured approach leads to increased student performance. Evaluation of the Québec reform brings further evidence and has, as mentioned above, crucial methodological advantages. First, the reform was mandated to each and every school across the province by the Ministry of Education, such that every school and teacher had to embrace the reform (at least to some degree) and students from all backgrounds were treated. The impact of the reform can be estimated on students with different abilities. Second, since teaching in the RofC continued to be delivered in the same way throughout the period, students from the other provinces may be used as controls. Third, the reform combined with our data set allows us to observe students treated for up to nine years (as opposed to three and less in the studies cited above). This longer observation period also allows us to estimate (to some extent) the long term effects of the reform.[15] Fourth, although the treatment period is long, because the reform was implemented in steps, all of our estimates are derived from a comparison of students in the same grade close by in time, which reduces the possible bias induced by other reforms taking place at the same time.

3. Empirical strategy

In economics, difference-in-differences (DID) methods have often been used to estimate the effects of policy reforms. Angrist and Krueger (1999) and Blundell and Costa Dias (2009) describe applications and give an

[14] The estimated effects are about one tenth to one seventh of a standard deviation.

[15] Borman et al. (2003) found that greater impacts were estimated when the school reform evaluated had been in place for a greater number of years (more than 5 years).

overview of the methodology. This approach can be used in settings where some individuals of a population are subject to a policy reform (or a treatment) while others are not, and comparable groups of individuals are observed prior to the policy intervention. The standard DID has raised a number of concerns in the literature (e.g. Bertrand, Duflo, & Mullainathan, 2004; Besley & Case, 2000; Donald & Lang, 2007). As a result, in addition to standard DID, we also use the changes-in-changes (CIC) model developed by Athey and Imbens (2006).

The CIC model relaxes some of the assumptions of the standard DID (the standard DID is a special case of the CIC model). Standard DID assumes outcomes are additive in time period, group and unobservable characteristics of the individual, while the CIC model is nonparametrically identified. Standard DID often assumes that the treatment effect is constant across individuals, or more generally assumes that the effect might differ across individuals but that the distribution of outcomes without treatment is common across groups. In the CIC approach, the distribution of the unobservable characteristics of individuals may differ across groups and the treatment effect may also differ according to the unobservable characteristics of the individual. In contrast to DID, the more general CIC model can accommodate the possibility that treated individuals may benefit more from the treatment than untreated individuals, and that the policy may have been implemented because greater benefits were expected in the treatment group (see Besley and Case (2000) on endogenous policy reform).

Using CIC, one can estimate the entire counterfactual distribution of outcomes in the absence of treatment for treated individuals, and in the presence of treatment for non-treated individuals. It is thus possible to evaluate the mean effect of a policy intervention, and also the effect at different points on the distribution. Quantile DID, which applies DID to each quantile as opposed to the mean, can also look into the distributional effects of the treatment. While individuals are compared across time according to their quantile in both QDID and CIC, they are compared according to their outcome in CIC and their quantile again in QDID.[16] QDID assumes that the underlying distribution of unobservable characteristics of the individuals is the same in both groups, while CIC allows for heterogeneity across groups. Athey and Imbens (2006) show the advantages of CIC over QDID. This feature is particularly attractive in the present application, as knowing whether lower performing students benefited more or less, compared to middle to top performing students, is of great policy interest.[17]

The CIC model relies on two main assumptions. First, the underlying production function for treated individuals and non-treated individuals, mapping the relationship between the outcomes and the unobservable characteristics at a given point in time, does not vary across groups. Second, the distribution of the unobservable component of the math score should remain the same within a group over time. As long as these hold true, CIC provides consistent estimates of the effect of treatment on both treated and untreated individuals.

In sum, CIC allows the possibility of time and treatment effect heterogeneity. It accommodates the possibility of selection into treatment due to expected larger benefits from treatment. It provides consistent estimates of the entire counterfactual distribution of outcomes of treated and non-treated individuals, and allows the two distributions to differ. As such, CIC permits policy evaluations in terms of a mean-variance trade-off.

In our setting, repeated cross-sections of students are observed in a treatment and a control group, before and after the treatment. Each child i is observed once, in time period $T_i \in \{0, 1\}$, where period 0 is prior to the school reform and period 1 is after the implementation of the school reform. Each student i also belongs to a group, $G_i \in \{0, 1\}$, where group 0 is the RofC (the control group) and group 1 is Québec (the treatment group). Effectively, we implement the following CIC estimator:

$$\tau^{CIC} \equiv E[Y^I_{11}] - E[Y^N_{11}] = E[Y^I_{11}] - E[F^{-1}_{Y,01}(F_{Y,00}(Y_{10}))], \qquad (1)$$

where Y^I_{gt} is the outcome of students receiving treatment in group g in time period t and Y^N_{gt} is the outcome of students not receiving treatment in group g in time period t. In equation 1, Y^I_{11} is the outcome of students receiving treatment in Québec after the implementation of the school reform and Y^N_{11} is the outcome of students not receiving treatment in Québec after the implementation of the school reform. Y^N_{11} is not observed, but can be inferred using $E[F^{-1}_{Y,01}(F_{Y,00}(Y_{10}))]$. $F_{Y,01}$ and $F_{Y,00}$ are the outcome distribution functions $F_{Y,gt}$ of students in the RofC after and before the implementation of the school reform respectively, and Y_{10} is the outcome of students in Québec prior to the reform. Since the math score in a given grade takes on twenty different values, we implement the discrete CIC model with conditional independence, and also provide the discrete CIC model lower and upper bounds.

The main identification condition for the estimation of the reform effect is that, aside from the Québec school reform, there are no other province-specific transitory shocks during the period pertinent to the performance of students in mathematics.[18] While we recognize that outcome variables related to the labor market (e.g. earnings, probability of employment) are subject to aggregate transitory shocks impacting the business cycle, we argue this is unlikely to be the case for test scores. The robustness of our results to this condition is further discussed in Section 5. Identification further requires that students did not self-select into or out of the reform (i.e. there were no compositional changes). Since the reform was implemented across all schools in Québec, parents were not able to self-select their children into (or out of) the reform. Furthermore, mobility between Québec and

[16] For recent evaluations of education policies focusing on quantile distributional effects see Jackson and Page (2013), Ryan (2013).

[17] Gender effects would also be of policy interest, but our sample size does not allow subgroup analysis.

[18] Bertrand et al. (2004) and Donald and Lang (2007) raise concerns related to the computation of standard errors. As pointed out by Athey and Imbens, their proposed solution relies on linearity and additivity.

the RofC, following a school reform, is extremely unlikely. As such, we can safely assume that students did not self-select into or out of the reform.

Deke and Haimson (2006) show that some students are more likely to benefit from an improvement in academic competencies, and that the gains are greater the weaker the student is in this area. In this spirit, one could expect the reform to have a greater impact on students weaker in mathematics than on highly performing students. As mentioned above, CIC allows us to estimate the impact of the reform at different points on the skill distribution. More specifically, the counterfactual distribution is obtained using the observed distributions of outcomes of the treated before treatment, the controls before treatment and the controls after treatment. In our setting, each point on the distribution is inferred as follows. First, treated students before treatment with a given score corresponding to a certain percentile, are associated with control students before the reform with the same score, but possibly located at a different percentile in the score distribution. Second, these control students prior to the reform are associated with control students after the reform located at the same percentile on the score distribution. The control students' change in score post reform is the inferred change in score of the treated students post reform had they not been treated located at the same percentile as treated students prior to the reform.[19] Comparing the score distribution of the treated individuals after treatment with the counterfactual distribution of the treated individuals had they not received treatment at different points on the distribution allows us to estimate the impact of the reform for lower skilled students as well as highly skilled students.

Individual characteristics (denoted X) need not be stable over time or across subpopulations, as long as the changing characteristics are observed. In our approach to CIC, we control for X through a linear specification. Standard DID also assumes that Y_{gt} is linear in X, such that the estimated response to the reform is also linear in X. To address the possibility of non linearity of response with respect to X, we also implement the matching difference-in-differences (MDID) (Heckman, Ichimura, & Todd, 1997, 1998). This estimator also allows the possibility of selection into treatment.[20] With repeated cross-sections of data, the MDID estimator is (Blundell & Costa Dias, 2009):

$$\tau^{MDID} = \sum_{i \in S_{11}} \left\{ \left[y_{it_1} - \sum_{j \in S_{10}} \tilde{w}_{ijt_0} y_{jt_0} \right] - \left[\sum_{j \in S_{01}} \tilde{w}_{ijt_1} y_{jt_1} - \sum_{j \in S_{00}} \tilde{w}_{ijt_0} y_{jt_0} \right] \right\} w_i \qquad (2)$$

where student i is in the treatment group after the reform and student j is part of a subpopulation S_{gt}, and can either be part of the treatment group prior to the reform S_{10}, the control group prior to the reform S_{00} or the control group after the reform S_{01}. The outcome variables are measured at time t_0 (prior to the reform) for students in S_{10} and S_{00}. The outcome variables are measured at time t_1 (after the reform) for students in S_{11} and S_{01}. Each student j when compared to student i is attributed a new weight $\tilde{w}_{ijt}$ that depends on the matching technique used. The sampling weight of student i is denoted w_i. The MDID estimator controls for X semi-parametrically by ensuring that students in each of the four groups are as close as possible to each other in terms of the distribution of X.

Effectively, we first estimate a probit model in which the dependent variable equals one if the student lives in Québec and equals zero otherwise. Using this model, we predict the propensity score of each student and perform matching using these scores. We implement kernel matching, local linear regression matching and nearest neighbor matching. Bootstrap standard errors are calculated to account for the underlying matching procedure. Rosenbaum and Rubin (1983) show that if observations in the treated and control groups have the same propensity score distribution, the underlying characteristics used to calculate the propensity score are also distributed equally. We include the following covariates: maternal education dummies, gender, area of residence, household income quartile, and a maternal work dummy. To assess the importance of non-parametrically controlling for X, we compare the estimated impacts using standard DID with those of MDID.

4. Data set

The data set used for our empirical analysis is Statistics Canada's National Longitudinal Survey of Children and Youth (NLSCY),[21] a long-term biennial survey designed to provide information about the development and well-being of Canadian children and youth. The survey covers a comprehensive range of topics including childcare, schooling, physical development, cognitive skills and behavior of the child as well as data on demographics and the social environment of the child (family, friends, schools and community). The NLSCY began in 1994–1995 (wave 1), and the last collection period to have been released by Statistics Canada covered 2008–2009 (wave 8). The sampling unit is the child (or youth). The NLSCY is designed to provide estimates representative of the provincial population of Canadian children aged 0–11 years old, first selected in wave 1 (1994–1995) of the survey.[22] For simplicity, from here on, we refer to the first year only to identify data in a particular wave. For example, data from wave 1 will be referred to as data from 1994.

The NLSCY contains three cohorts of longitudinal children. The main cohort includes children aged 0–11 in 1994. At the time the NLSCY survey was last conducted

[19] See Figure 1 in Athey and Imbens (2006) for a graphical representation.

[20] Variations in response rates across survey waves may create a selection bias.

[21] Many studies in economics have used this data set (e.g. Baker, Gruber, & Milligan, 2008).

[22] Weights, adjusted for total non-response matching known population counts, are provided in each wave of the survey.

(year 2008) these children were aged 14–25. This longitudinal sample is central to our study as it provides information on primary and secondary school students from academic years 1994–95 (grades 1–6 or 6–11 year olds) to 2008–09 (grades 9–11 or 14–17 year olds). The other two cohorts provide additional observations on primary school students in grades 1–4 in academic year 2006–07, and in grades 1 and 2 in academic year 2008–09.

In sum, the NLSCY provides three cohorts of children of primary and secondary school age: (1) students in grades 1–6 in academic year 1994–95 up to grades 9–11 in academic year 2008–09, (2) students in grades 1–4 in academic year 2006–07, and (3) students in grades 1 and 2 in academic year 2008–09.

The next four subsections are organized as follow. First, we describe the measure of mathematical ability used in this study and relate it to later outcomes, such as total personal income. Second, we present the school grades for which we observed students passing comparable math tests, by year and by treatment status. Third, we show the summary statistics for these students by treatment group (Québec vs RofC). Fourth, we provide an overview of the trends in math scores by treatment group, and also discuss the number of years of treatment and the response rate to the test for each subgroup.

4.1. Measure of mathematical ability

The NLSCY provides one measure of achievement in mathematics for school age children: the CAT/2 mathematics test.[23] The CAT/2 test is a shorter version of the Mathematics Computation Test taken from the Canadian Achievement Tests, 2nd edition. This test was developed by the Canadian Test Center after careful consideration of the differences among the main school curricula across Canada. The CAT/2 is designed to measure procedural skills in mathematics. The test consists of 20 questions on four binary operations (i.e. addition, subtraction, multiplication, and division) on integers, decimals, fractions, negatives, exponents and percentages, and on simple linear algebra. The test is administered to students enrolled in grades 2–10, aged 7–15. The difficulty of the test varies with the school grade of the child. Thus, there are different tests depending on the school level of the child. Grade 2 students passed the level 2 test, grade 3 students the level 3 test, and so on. The master files of the NLSCY provide both the raw scores and the standardized scores. The raw score is simply the number of correct answers to the test. The standardized scores are obtained using sub-samples (by schooling-grade) of the normative sample.[24] The standardized scores are designed to numerically represent the relative level of mathematics a child has attained and to track the progress of a child in mathematics throughout the years. The range of scores for grade 2 students is thus much lower than the range for grade 10 students, but overlaps with the range of grade 3 students such that particularly strong 2nd graders may be as proficient in mathematics as some lower performing 3rd graders. Since the level depends on the school grade and not the age of the child, it is possible that students of different ages passed the same test in a particular survey wave, and that students of the same age passed different tests.

The CAT/2 test is comparable to the test used in Murnane et al. (1995), where elementary mathematical concepts such as the working of fractions and decimals are measured using a multiple choice test of 25 questions. It is also aligned with the definition of mathematical development in Ingram and Neumann (2006) used to assess the importance of mathematics in different types of occupations. To further validate the relevance of the test used in this study, we estimate the correlation between the CAT/2 score and a variety of adult outcomes. Using the oldest cohort of the NLSCY, youth aged 24–25, we regress a high school dropout dummy, the highest level of education completed, and the total personal income on the CAT/2 score at age 12–13.[25] Results are shown in Table 1. If we assume that the marginal effects are constant, we find that the dropout rate decreases by 0.07 of one percentage point if the math score increases by 1, the probability of obtaining a University degree increases by 0.12 of one percentage point and the total yearly personal income increases by 25$. We later quantify the effect of the reform in terms of these three adult outcomes in Section 5.1.

We restrict our attention to years in which comparable tests were administered.[26] Table 2 provides a more detailed overview of the grades and academic years observed using the NLSCY CAT/2 test. The table covers academic year 1996 to academic year 2008. Table 2 shows that students entering grade 1 in academic year 1989 or 1990 are later observed in the same school cycle (i.e. grades 7–8)[27] during the same academic year. This is also true for students entering grade 1 in academic year 1991 or 1992, and so on. Since the reform was designed by cycle, this grouping comes by naturally.[28]

[23] In 1996 and 1998, a reading test was also administered (in addition to the CAT/2 mathematics test). However, as of 2000, the reading test was discontinued because of time constraints.

[24] More specifically, Statistics Canada standardizes the raw scores using a sample of Canadian children from the ten provinces called "the normative sample". This sample received the complete Mathematics Computation Test.

[25] Youths aged 24–25 are observed in 2008 only, and approximatively 1400 observations are available. These young adults passed the CAT/2 test in 1994, 1996 and 1998, but we use the CAT/2 score ages 12–13 in 1996 (grades 5–6). Scores from 1994 are excluded because of the ceiling effect detected in the 1994 tests (see Web Appendix for further details). Scores from 1998 are excluded because of the low response rate (see Web Appendix for further details).

[26] Our Web Appendix provides further details on the comparability of the test scores over the years.

[27] As mentioned above, the reform in primary school was implemented by cycle: with cycle 1 comprised of grades 1 and 2, cycle 2 grades 3 and 4, and cycle 3 grades 5 and 6. We extend this grouping to secondary school and regroup grade 7 and 8 students together, and grade 9 and 10 students together.

[28] As previously mentioned, the entire school program was generally designed by cycle. In practice, some schools implemented multi-grade classrooms using the same grouping structure by cycle. Some schools also assigned teachers to one group of students for two years, such that students only had one teacher per cycle.

Table 1
Influence of math score on later outcomes.

	High school dropout	Highest level of education	Total personal income
CAT/2 score			
Coef.	−0.015120	0.009723	25.367860
dy/dx	−0.000665	0.001226	25.367860
z	−4.74	4.18	2.16
Regression	Logit	Ord. logit	Linear

Note: Shows the association between the math score and adulthood outcomes. We use observations from the last wave of the NLSCY since it provides us with the oldest cohort of children (age 24–25). We use the math score for which the number of respondents was the highest given the age group: grades 5–6 score (survey year 1996). We report the marginal effect (dy/dx) of the math score. For the highest level of education, we report the margin for the probability of having a University degree. Controls include gender, age, province of residence, area of residence and the education level of the primary care giver at the time of the CAT/2 test. Standard errors (not reported here) are bootstrapped to account for the clustering and stratifications of the NLSCY.

4.2. The incidence of the reform

Table 2 also provides information on the incidence of the reform: boxed grades are under the reform, while unboxed grades are prior to the reform. Our main treated sample of students entered grade 1 in 1999 or 2000. Students entering grade 1 in 1999 were treated as of grade 2 in academic year 2000, while students entering grade 1 in 2000 were treated from the start of grade school. This group constitutes our main longitudinal sample, as they are observed from grades 2–10 after the reform was implemented. It allows us to estimate the cumulative effect of the reform.

We also observe three subsidiary samples: students entering grade 1 in year 2003 and 2004, students entering grade 1 in 2005, and students entering grade 1 in 2007. The first sample is observed in grades 3–4, while the second and third samples are only observed in grade 2. Since these three samples contain observations on students entering the reform at a later point, estimates using these alternative samples allow us to further validate the results obtained from our main longitudinal sample. Students in these subsidiary samples were however impacted by the Québec childcare reform, but students in grades 3–4 were in childcare at a time (in 2000) when the number of spaces remained fairly constrained. Research suggests that the childcare reform may have had a negative impact on the cognitive development of children (Haeck, Lefebvre, & Merrigan, 2013). Our empirical strategy allows us to control for the specific impact of the childcare reform. Further discussion is presented in Section 5.1.

A special case concerns students that were only treated by the reform while in grade 4 of academic year 2001 (unboxed grades in bold with an asterisk "*"). These students are the first students observed in the NLSCY to be treated by the reform. They entered grade 1 in 1998 and were only treated while in grade 4 of academic year 2001. As such, this group is only partially treated and for a period of one year only. This allows us to assess the impact of the reform on older children when first implemented for one year only. Looking at their test scores later, we can also determine whether the effect persists or not.

4.3. Student and family characteristics

There is no reason to believe that students from Québec differ in a meaningful way from students in the RofC. Nonetheless, using the NLSCY we compare both groups of students on a number of student and family characteristics. In Table 3, the summary statistics of students attending school in Québec are presented in the first two columns and those of students attending school in the RofC in the last two columns. The NLSCY provides a total of 7745 observations for the Québec sample and 33,390 observations for the RofC for our purpose. The mean and standard deviations reported in Table 3 are weighted using the population weights provided by Statistics Canada.

Table 3 shows that an equal proportion of male and female students are observed in both groups. These students also have comparable scores on the Peabody Picture Vocabulary Test (PPVT) taken at ages 4 and 5. The PPVT is widely used in the literature related to early childhood cognitive development to assess receptive and hearing vocabulary and may serve as a measure of early childhood ability. This measure allows us to control for reforms impacting the early development of children, more specifically it allows us to account for the possible impact of the childcare reform for some of the subsidiary samples. The proportions of students per school cycle (grade 2, grades 3 and 4, and so on) are also comparable.

Students from Québec are similar to RofC students in terms of family characteristics. Most students live in a two-parent household and have a mother who works, but students in the RofC have mothers who are slightly more educated, their family income is generally higher and they are less likely to live in a highly populated urban area. Since the cost of living varies both by province and in time, in our empirical approach we use household income quartile by province and by year when controlling for confounders.

Although students observed in both groups share similar characteristics, we control for all of these characteristics in our empirical approach to ensure that our estimated effects are not a mere reflection of differential changes in X over time.

4.4. Math scores over time and grade levels

Math scores over time are presented in Table 4. This table provides a detailed overview of the test scores by subgroups over time.

Table 4 shows the mathematical assessment summary statistics for the different groups of treated versus non-treated students by school grade and academic year. The first panel shows the summary statistics for grade 2

Table 2
Reform schedule on students observed by school grade and academic year.

Grade 1 entry year	Academic Year						
	1996	1998	2000	2002	2004	2006	2008
1989–90	7 – 8	.					
1991–92	5 – 6	7 – 8	.				
1993–94	3 – 4	5 – 6	7 – 8	9 – 10			
1995–96	2	3 – 4	5 – 6	7 – 8	9 – 10		
1997–98		2	3 – 4	**5*** 6	**7*** 8	**9*** 10	
1999–00			2	3 – 4	5 – 6	7 – 8	9 – 10
2001–02				.	.	.	.
2003–04					.	3 – 4	.
2005						2	.
2007							2

Note: Shows the school grades pre and post reform in which students are observed and passed the CAT/2 test. Students entering grade 1 in 1989 and 1990 are grouped together, and so are students entering grade 1 in 1991 and 1992, and so on. This grouping conforms to the reform approach which groups students by school cycle (e.g. grade 1 and 2). Boxed grades are under the reform, while unboxed grades are not. Grades in bold with an asterisk "*" are not under the reform, but students in these grades were treated by the reform while in grade 4 in academic year 2001. An empty cell is denoted by ".". A cell is empty if the number of students observed is equal to zero or only includes a small number of grade repeaters, or if the test administered that year was not comparable to all other years (grades 9–10, academic year 1996–2000).

Table 3
Summary statistics.

	Québec (treated)		Rest of Canada (control)	
	Mean	Std. dev.	Mean	Std. dev.
Student characteristics				
Male	0.49	(0.50)	0.51	(0.50)
PPVT (age 4–5)	99.64	(15.11)	100.16	(14.81)
School grade				
2	0.10	(0.30)	0.11	(0.31)
3 and 4	0.22	(0.42)	0.22	(0.41)
5 and 6	0.24	(0.43)	0.23	(0.42)
7 and 8	0.28	(0.45)	0.27	(0.44)
9 and 10	0.16	(0.36)	0.18	(0.38)
Family characteristics				
Family structure				
One parent	0.21	(0.41)	0.18	(0.38)
Two parents	0.79	(0.41)	0.82	(0.39)
Maternal education				
Less than secondary	0.16	(0.37)	0.09	(0.29)
Secondary	0.24	(0.42)	0.23	(0.42)
Some post-secondary	0.18	(0.38)	0.20	(0.40)
College or university	0.41	(0.49)	0.47	(0.50)
Mother works (dummy)	0.81	(0.39)	0.84	(0.37)
Household income ('000s)	66.92	(47.26)	77.56	(60.44)
Area of residence				
Rural	0.14	(0.34)	0.13	(0.34)
Urban, $\leqslant$30,000	0.15	(0.36)	0.19	(0.39)
Urban, 30,000–99,999	0.09	(0.29)	0.09	(0.29)
Urban, 100,000–499,999	0.06	(0.24)	0.21	(0.41)
Urban, $\geqslant$500,000	0.55	(0.50)	0.38	(0.48)
No. of obs.	7745		33,390	

Note: Shows the mean and standard deviation for a number of student and family characteristics of Québec students (left) and RofC students (right). The sample is restricted to students' observations used to compute the estimated impact of the reform: all grade 2 students, students in grade 3–8 in academic year 2000–2006 (except for grade 7 and 8 students in academic year 2004), and students in grades 9 and 10 in academic year 2002, 2004 and 2008.

Table 4
Math scores summary statistics by school grade, academic year and treatment group.

		Québec (treated)							Rest of Canada (control)						
		1996	1998	2000	2002	2004	2006	2008	1996	1998	2000	2002	2004	2006	2008
Grade 2															
Years in reform		0	0	**1**	.	.	**2**	**2**	0	0	0	.	.	0	0
CAT/2	Mean	340	320	**298**	.	.	**284**	**301**	313	293	289	.	.	279	280
	Std. dev.	(48)	(40)	**(39)**	.	.	**(38)**	**(39)**	(47)	(38)	(42)	.	.	(40)	(39)
	% response	65	49	**91**	.	.	**94**	**94**	77	52	75	.	.	88	86
	No. obs	135	114	**285**	.	.	**242**	**163**	683	506	929	.	.	1065	906
Grades 3–4															
Years in reform		0	0	0	**3**	.	**3–4**	.	0	0	0	0	.	0	.
CAT/2	Mean	431	402	391	**377**	.	**353**	.	394	375	358	366	.	347	.
	Std. dev.	(60)	(48)	(53)	**(44)**	.	**(53)**	.	(58)	(55)	(52)	(49)	.	(52)	.
	% response	70	40	90	**91**	.	**95**	.	74	51	81	84	.	88	.
	No. obs.	281	185	396	**572**	.	**787**	.	1377	1076	1429	2022	.	3862	.
Grades 5–6															
Years in reform		0	0	0	1-0	**5**	.	.	0	0	0	0	0	.	.
CAT/2	Mean	507	484	462	465	**436**	.	.	474	451	431	440	436	.	.
	Std. dev.	(48)	(50)	(53)	(53)	**(55)**	.	.	(60)	(59)	(56)	(51)	(53)	.	.
	% response	73	47	90	95	**94**	.	.	73	53	77	88	89	.	.
	No. obs.	292	184	307	388	**528**	.	.	1322	998	1275	1469	1956	.	.
Grades 7–8															
Years in reform		0	0	0	0	1-0	**7**	.	0	0	0	0	0	0	.
CAT/2	Mean	590	533	540	523	517	**495**	.	540	504	492	485	473	485	.
	Std. dev.	(73)	(57)	(71)	(58)	(60)	**(67)**	.	(78)	(69)	(73)	(68)	(60)	(70)	.
	% response	73	41	89	93	94	**91**	.	76	45	76	77	89	87	.
	No. obs.	225	175	269	336	354	**518**	.	1241	878	1147	1236	1402	1834	.
Grades 9–10															
Years in reform		.	.	.	0	0	1-0	**9**	.	.	.	0	0	0	0
CAT/2	Mean	.	.	.	637	609	605	**599**	.	.	.	583	579	567	596
	Std. dev.	.	.	.	(90)	(82)	(87)	**(86)**	.	.	.	(85)	(91)	(81)	(88)
	% response	.	.	.	87	92	91	**84**	.	.	.	66	84	86	82
	No. obs.	.	.	.	184	221	261	**343**	.	.	.	864	1166	1209	1538

Note: Shows the summary statistics on the CAT/2 test (mean, standard deviation, response rate to the test and number of observations) by school grade, academic year and treatment group (Québec vs RofC). Summary statistics on treated students post reform are in bold. The response rate is the ratio of the weighted number of students with non missing scores on the CAT/2 test over the total weighted number of students. Years in the reform refer to the number of years of treatment (it is always equal to zero in the RofC, while it is only equal to zero for Québec students not treated by the reform). The number of years in the reform for grade 3–4 students in 2006 is "3–4" since it is equal to 3 for students in grade 3 and equal to 4 for students in grade 4. The same logic holds for students in grades 5–6 in 2002, 7–8 in 2004 and 9–10 in 2006. An empty cell is denoted using a "." (see Table 2's footnote for further details).

students, the second panel, grades 3–4, the third, grades 5–6, the fourth, grades 7–8, and the fifth, and last panel grades 9–10. The left panel shows the summary statistics for students residing in Québec (treatment group), while the right panel includes only students residing in the RofC (control group). For each grade (or combined grades) we show first the number of years of treatment (always equal to zero in the RofC as the reform was implemented in Québec only). This allows the reader to better understand which groups can be compared and what is the intensity of treatment for each group. Second, we show the summary statistics related to the CAT/2 test: first the mean score, second the standard deviation, third the number of observations, and fourth the response rate to the test. The response rate is calculated using the population weight and thus provides a representative percentage of students in the population who completed the test.

From grade 2 to grades 9–10, we find that the mean score is increasing in both groups each year. As mentioned above, the CAT/2 test is designed to numerically represent the progression of students in mathematics throughout the years. Comparing Québec and the RofC year by year, we find that Québec students consistently score higher than comparable students in the RofC,[29] but trends over time are different.

In grade 2 (top panel), given the response rates, we decide to exclude 1998 observations from our main analysis. The response rate in 1998 is 49% in Québec and 52% in the RofC. As a result, for grade 2, we use 1996 as the base year (prior to the reform), and 2000, 2006, and 2008 as the treatment years.[30] In 1996, the response rate was below 80%, while it was generally above 90% in 2000, 2006, and 2008. In 1996, the tests were being administered in schools at the end of the school year. If schools with lower performing students were more likely to not administer the tests due to time constraints days before the final exams, then mean score values in 1996 are overestimated, in both Québec and the RofC. If they are

[29] One exception is grades 5–6 in 2004 where the test scores are equal in both regions.

[30] Grade 2 students are not observed in 2002 and 2004 in the NLSCY (see Table 2). Results using 1998 are presented in the Appendix.

overestimated by the same magnitude, our estimates should be unbiased. Nonetheless, we specifically address the possibility of a bias due to the variation in non response in the empirical section. The summary statistics over time suggest that the scores have been downward trending in both groups (Québec and RofC), but the decrease has been more striking in Québec.

In both grades 3–4 and grades 5–6, we use 2000 as the base year prior to the reform given the lower response rate in 1996 and 1998 in both groups. Looking at the results between 2000 and 2006, we find that mean scores are decreasing in Québec post reform (for the academic year in which the number of years in the reform is greater than zero for all students),[31] while no clear pattern can be identified in the RofC. Focusing again on results from 2000 and beyond, the fourth panel on grade 7–8 students reveals that the mean scores in Québec prereform (before 2004) were largely above the post reform (in 2006) results, while they were generally stable in the RofC, decreasing slightly in 2004. Finally, for grade 9–10 students (last panel), the mean values suggest an important decrease in Québec when comparing 2002 and 2008 outcomes, and a more modest decrease for 2004 and 2006 versus 2008. Grade 9–10 students in the RofC generally perform better in 2008 (compared to 2002, 2004 and 2006).

Focusing on years when the response rate is high, students in Québec generally perform better than students in the RofC across all grades and across all time periods, but the mean score generally decreases more sharply in Québec post reform in all grades, while the pattern in the RofC is generally stable (or at least shows no precise trend). Section 5 further validates these results by computing the statistical significance of those differences using standard DID and MDID. Distributional (and mean) effects are computed using the CIC approach and the possibility of pre-existing trends[32] is specifically accounted for in Section 5.2.

5. Effects of the reform on math scores

Table 5 presents the empirical results using DID and MDID. Estimated impacts using DID and MDID with three matching techniques, suggest that the reform had significant negative effects on mathematical abilities. Given the complex sampling design of the NLSCY, all estimations are performed using the sampling weights provided by Statistics Canada. To account for the clustering and stratification of the NLSCY, the standard errors are estimated using the 1000 bootstrap weights provided by Statistics Canada. We first present the results of our benchmark specification using standard DID. Then we assess the robustness of our results to students' heterogeneity, non response, other reforms in the RofC, and the linearity assumption using MDID, and we present a number of falsification exercises to further validate our results. We then confirm the results from standard DID using CIC and discuss the distributional effects. Finally, we discuss the materiality of the estimated impacts and verify the robustness of our results when allowing for a pre-existing trend in the outcome variable.

5.1. Benchmark specifications

We first focus on results presented in the first two columns of Table 5. The first specification (column 1) does not control for any covariates, while the second specification (column 2) controls for gender, school grade, maternal education, household income quartile, a dummy indicating whether the mother works or not and the area of residence.[33] Across all grades, the estimated impacts of the reform are negative and statistically significant. For the first specification, in grade 2, the estimated impacts, all negative, range from 6.1 to 22.3 (1.8–6.6% of the mean score prereform). In grades 3–4, they range from 22.0 to 27.6 (5.6–7.1% of the mean score), while in grades 5–6 they range from 13.4 to 20.3 (2.9–4.3% of the mean score). Finally, in grades 7–8, the estimated effects range from 22.5 to 36.9 (4.3–6.8% of the mean score) and in grades 9–10 they range from 24.0 to 51.5 (4.0–8.1% of the mean score).

Results from the second specification (column 2), DID with covariates, are generally comparable: 11.1–24.3 (grade 2), 16.7–19.6 (grades 3–4), 13.4–20.1 (grades 5–6), 23.1–33.8 (grades 7–8), and 28.6–45.1 (grades 9–10). This suggests that the magnitude of the estimated effects are larger the higher the school grade (both in absolute value and in % of the mean score prereform), but are of comparable magnitude from grade 7 to 10 (in units of a standard deviation). As students in higher grades have been exposed to the reform for a longer period, this finding suggests that the reform consistently limits the development of students in mathematics compared to the pre-reform approach. One exception are students in grades 5–6 in academic year 2002. Students in grade 5 had only been in the reform for one year (in grade 4 in 2001) and students in grade 6 had not been in the reform (see years in Table 4). It is therefore not surprising to find that the estimated effect is smaller in magnitude for this cohort. More surprising are the large effects estimated for grade 2 students (of about 25–50% of a standard deviation).

5.1.1. Students' heterogeneity

The estimated effects of the reform reported above do not account for students' unobserved heterogeneity. We do not have any reason to believe that the distributions of students' unobserved ability differ across provinces (or more specifically between Québec and the RofC), except for the later cohorts (grade 2 years 2006–2008, grades 3–4

[31] As mentioned above, students in grades 5–6 in 2002 were only partially treated. The number of years for this group is denoted "1-0" because grade 5 students were treated for one year, while in grade 4, and grade 6 students were never treated.

[32] Trends in mean math scores over time are further discussed in the Appendix.

[33] Results are robust to the inclusion of age in months at the time of test, and province of residence. Because entry age regulations are different across provinces, the support for the age at the time of test in Québec is different from that of the RofC. Since age and province do not alter our results but cannot be included when estimating MDID (common support issue), we generally do not include these variables.

Table 5
Estimated effects of the treatment on the treated.

Dependent variable: CAT/2 score	Base specifications		Alternative specifications					
	DID		DID			MDID		
	Without covariates	With covariates	With PPVT	Impute m(10)	Excl. Ontario	Kernel	llr	nn (5)
Grade 2								
Years 1996, 2000								
Coef.	−17.32**	−21.18***	−14.23	−15.64**	−21.04**	−17.86***	−18.47**	−16.76**
Std. err.	(8.38)	(8.17)	(9.27)	(7.88)	(9.01)	(5.91)	(8.50)	(7.79)
No. obs.	2032		1668	2620	1502		1139	
Years 1996, 2006								
Coef.	−22.29***	−24.31***	−19.89**	−18.51**	−32.70***	−26.83***	−25.67***	−18.86**
Std. err.	(8.37)	(8.09)	(9.33)	(7.97)	(8.57)	(5.33)	(8.02)	(8.28)
No. obs.	2125		1809	2543	1510		968	
Years 1996, 2008								
Coef.	−6.06	−11.07	−5.13	−6.05	−11.51	−7.34	−7.00	−2.52
Std. err.	(8.55)	(8.37)	(10.04)	(8.09)	(8.86)	(6.50)	(10.34)	(10.05)
No. obs.	1887		1565	2317	1312		652	
Grades 3–4								
Years 2000, 2002								
Coef.	−21.97***	−19.64***	−16.80***	−17.40***	−16.21***	−11.06***	−10.39**	−6.41
Std. err.	(5.03)	(4.43)	(4.50)	(4.56)	(4.73)	(3.67)	(4.69)	(4.52)
No. obs.	4419		4088	5194	3303		2278	
Years 2000, 2006								
Coef.	−27.60***	−16.66***	−14.12***	−15.13***	−18.03***	−11.04**	−11.78**	−10.01*
Std. err.	(5.27)	(4.87)	(4.91)	(4.94)	(5.30)	(4.45)	(5.42)	(5.16)
No. obs.	6474		6115	7818	4943		3133	
Grades 5–6								
Years 2000, 2002								
Coef.	−13.40**	−13.43**	−12.54*	−12.69**	−9.41	−14.37***	−13.92**	−16.57***
Std. err.	(6.23)	(5.97)	(6.63)	(6.09)	(6.13)	(4.94)	(5.99)	(5.96)
No. obs.	3439		3139	4053	2608		1550	
Years 2000, 2004								
Coef.	−20.30***	−20.13***	−17.79***	−18.33***	−19.41***	−21.64***	−22.22***	−17.21***
Std. err.	(6.12)	(5.86)	(6.50)	(6.25)	(6.12)	(4.80)	(6.24)	(6.66)
No. obs.	4 066		3632	4698	3054		2103	

Table 5 (*Continued*)

Dependent variable: CAT/2 score	Base specifications		Alternative specifications					
	DID		DID			MDID		
	Without covariates	With covariates	With PPVT	Impute m(10)	Excl. Ontario	Kernel	llr	nn (5)
Grades 7–8								
Years 2000, 2006								
Coef.	−36.86***	−33.43***	.	−29.85***	−31.07***	−32.06***	−34.58***	−30.57***
Std. err.	(7.81)	(7.46)	.	(7.70)	(7.71)	(7.00)	(8.31)	(8.32)
No. obs.	3768		.	4446	2817		2059	
Years 2002, 2006								
Coef.	−22.47***	−23.08***	−26.44***	−21.86***	−27.70***	−30.11***	−34.37***	−27.82***
Std. err.	(7.52)	(6.70)	(6.49)	(7.10)	(7.25)	(6.38)	(8.31)	(7.99)
No. obs.	3924		3471	4599	2944		2067	
Years 2004, 2006								
Coef.	−31.25***	−33.89***	−32.40***	−30.66***	−38.44***	−24.04***	−21.27***	−14.77**
Std. err.	(7.49)	(6.74)	(6.94)	(7.14)	(7.40)	(6.38)	(7.38)	(7.37)
No. obs.	4108		3803	4627	3059		2065	
Grades 9–10								
Years 2002, 2008								
Coef.	−51.53***	−45.12***	.	−40.49***	−31.58***	−33.12***	−39.99*	−32.51***
Std. err.	(12.41)	(10.75)	.	(10.76)	(11.27)	(9.56)	(23.61)	(11.89)
No. obs.	2929		.	3719	2102		1364	
Years 2004, 2008								
Coef.	−23.98**	−28.58***	−29.30***	−24.04**	−25.30**	−39.11***	−34.73**	−30.05**
Std. err.	(11.47)	(10.73)	(11.24)	(10.42)	(11.56)	(9.95)	(16.47)	(12.44)
No. obs.	3268		2948	3832	2347		1360	
Years 2006, 2008								
Coef.	−28.63**	−33.75***	−39.44***	−28.14**	−18.72	−22.46**	−19.27	−17.29
Std. err.	(12.08)	(11.23)	(11.58)	(11.21)	(11.71)	(9.53)	(13.01)	(11.46)
No. obs.	3351		3093	3938	2404		1363	

Note: Shows the estimated effects of the reform by school grade (top panel grade 2, middle panel grades 3–4, bottom panel grades 5–6). Refer to Table 5's note on the next page for further details. *Note*: This table shows the estimated effects of the reform by school grade (top panel grades 7–8, bottom panel grades 9–10). Column 1 shows the effects using DID without covariates. Column 2 shows the effects using DID with covariates (i.e. gender, school grade, family structure, maternal education, household income quartile, maternal work, and area of residence). Models presented in columns 3–8 augment the model presented in column 2. In column 3, the PPVT score at age 4 and 5 is controlled for. Cells related to students first interviewed beyond age 4 and 5 are left empty and are denoted using ".". In column 4, missing scores are imputed using multiple imputation by chained equations. Estimated results are based on 10 imputations. In column 5, all students from Ontario (Canada's largest province) are excluded. Columns 6, 7 and 8 show the effects using MDID with kernel, local linear regression and nearest neighbor with 5 neighbors matching respectively. All standard errors are bootstrapped to account for the clustering and stratifications of the NLSCY. Coefficient significance is denoted using asterisks: $^{***}p < 0.01$, $^{**}p < 0.05$, and $^{*}p < 0.1$.

year 2006) impacted by the childcare reform. However, given the relatively small sample size used in this study compared to the overall Canadian population of students, we address this possibility for all of our estimates. To control for unobserved student heterogeneity, we use a widely used measure of early childhood cognitive development, the PPVT score. This score is completely unaffected by the Québec school reform, since students complete the PPVT at age 4 and 5 prior to entering primary school (i.e. grade 1).

Estimated impacts are reported in the third column of Table 5. This specification controls for the same set of covariates as specification 2, but also controls for the PPVT score. Comparing the number of observations in columns 1 and 3, we observe that the PPVT is generally available for about 82–94% of the students in our sample depending on the groups being compared. However, for students in grades 7–8 in 2000 (observed again in grades 9–10 in 2002), the PPVT score is completely missing. These students never took the test as they were first surveyed at ages 6 and 7 (in 1994). Estimated effects controlling for the PPVT scores (column 3) are slightly smaller in magnitude for grade 2 students. This is also true in grades 3–4 and 5–6, but in grades 7–8 and 9–10, the estimated impacts are slightly larger. This reinforces our earlier statement on the growing effect of the reform with increased exposure.

5.1.2. Non-response bias

Students' unobserved heterogeneity is not the only source of potential bias. The pattern of non response documented in Table 4 may also influence our results. So far, we have assumed that if scores were overestimated (or underestimated) they were overestimated (or underestimated) by the same magnitude in both groups. This may not be true, especially in cases where the non-response rate did not change in one group over time, but did in the other (e.g. grade 2, years 1996 versus 2000).

Since changing patterns in non response may influence the results, we imputed the missing math scores using multiple imputation by chained equations, and re-estimated the impact of the reform using the same covariates. The following variables were included in the imputation procedure: gender, school grade, maternal education, household income quartile, a dummy indicating whether the mother works or not, area of residence, province and year. The scores were imputed a total of 10 times ($M = 10$). The estimated coefficient $\bar{b}$ was obtained using:

$$\bar{b} = \frac{\sum_1^M b_m}{M},$$

where b_m is the regression coefficient of the m imputed data set (Schafer, 1997). The standard error was obtained using:

$$SE_{\bar{b}} = sqrt\left(\frac{\sum_1^M (SE_{b_m})^2}{M} + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_1^M (b_m - \bar{b})^2\right),$$

where SE_{b_m} is the standard error of the regression coefficient of the imputed data set m. Results are reported in the fourth column of Table 5. Across all grades, estimated effects $\bar{b}$ are generally slightly smaller, but again the confidence intervals overlap with that of the DID estimates with covariates (column 2).

5.1.3. Reforms in the rest of Canada

In Canada, primary and secondary school education is within provincial jurisdiction. Other school reforms taking place in other provinces could bias our results. So far we have assumed that teaching in the RofC continued to be delivered in the same way.

Our review of events around the time of the implementation of the Québec school reform does not suggest that other major reforms took place in the RofC, except in Ontario (Canada's largest province). In a recent report on the world's most improved school systems (Mourshed, Chijioke, & Barker, 2010), the province of Ontario is identified has having undertaken a whole system school reform (started in 2003), and having registered significant and sustained student outcome gains (based on 2003–2009 assessment data). The reform, mainly in the form of additional funds allocated to schools, encouraged schools to set their own objectives and to decide how to best address the needs of the least advantaged students. Schools were, for example, allowed to increase the number of working hours, to reduce the number of students per class, and to obtain help from education specialists.

Our reading of the evolution of Ontario's math scores on the CAT/2 test and PISA scores on comparable assessments (reading, math and science) between 2000 and 2009 does not suggest any progress (results are generally flat), but the Trends in International Mathematics and Science Study (TIMSS) results in science (but not math) suggest a potential gain. Nonetheless, we decided to exclude the scores for Ontario's students to assess the robustness of our results. Estimated impacts excluding Ontario are reported in the fifth column of Table 5. Estimated effects are some times slightly above that of our benchmark specification (column 2), and sometimes slightly below. In all cases, confidence intervals strongly overlap.

5.1.4. Linearity assumption

All of the estimated impacts presented above rely on the assumption that the *X*s' influence the outcome linearly. To ensure that our results are not dependant on this assumption, we also implemented MDID. Comparing the MDID (columns 6–8, Table 5) estimates with the DID estimates (column 2, Table 5), we find that the MDID estimates generally have confidence intervals that considerably overlap with those of the DID estimates using all three techniques (at 5%). Two main differences are noteworthy. First, in grades 3–4, the MDID estimators are smaller in magnitude and some times not significant. Second, in grades 9–10, for academic year 2006 compared to 2008, some of the MDID estimates are not significant, while the DID estimates are.

5.1.5. Falsification exercises

Finally, to further validate our results, we ran a number of falsification tests. Table 6 presents these results. We estimate the effects on all possible combinations of groups that are exclusively pre-reform (labeled "before" in the last

Table 6
Falsification exercises.

Dep. Var.: CAT/2	Coef.	Std. err.	No. obs.	All groups
Grade 2				
Years				
2000–2006	−11.02*	(5.65)	1788	After
2000–2008	9.49	(6.19)	1590	After
2006–2008	21.22***	(5.93)	1598	After
Grades 3–4				
Years				
1996–2000	−3.50	(5.74)	2642	Before
2002–2006	−2.02	(4.61)	5480	After
Grades 5–6				
Years				
1996–2000	1.41	(6.55)	2443	Before
2002–2004	−8.75*	(5.32)	3248	After[#]
Grades 7–8				
Years				
2000–2002	−1.61	(7.39)	2281	Before
2002–2004	9.73	(7.91)	2523	Before
2000–2004	6.09	(8.64)	2396	Before
Grades 9–10				
Years				
2002–2004	−8.93	(12.28)	1811	Before
2004–2006	−5.75	(13.28)	2113	Before
2002–2006	−10.86	(12.16)	1868	Before

Note: Shows the mean effect using standard DID with covariates (see Table 5's note for list of covariates) using groups that are either all before the reform or all after the reform (column 4, before vs after) in both Québec and the RofC. The dependent variable is the CAT/2 score. In grades 5–6, years 2002 versus 2004 are classified as after[#] because only grade 5 students were treated and for only one year while in grade 4 in 2001. Standard errors are bootstrapped to account for the clustering and stratifications of the NLSCY. Coefficient significance is denoted using asterisks: ***$p < 0.01$, **$p < 0.05$, and *$p < 0.1$.

column), and groups that are exclusively post reform (labeled "after" in the last column). Generally, we find no significant effects on all group combinations, but a few exceptions are worth discussing.

In grade 2, the effect is positive between 2006 and 2008. This relates to our earlier discussion on a possible neutral effect in the long run, as teachers further react to the reform by either mastering the reform content better, or by putting aside some of the reform content in favor of other teaching approaches. In grades 5–6, there is a weak negative effect when comparing 2002 and 2004 scores. As mentioned above, grade 5–6 students in 2002, were only partially treated (grade 5 students only, while in grade 4). As such, it is not surprising to find a negative effect. Overall, the falsification exercises support our conclusions.

In sum, the mean effects documented so far suggest that the reform had negative effects and that these effects are larger the longer a student was treated by the reform. These findings are in line with Crawford and Snider (2000) on the impact of a more academic approach against a more contextual approach on math scores. Results from our benchmark specification are robust to a variety of alternatives, but all rely on the more restrictive assumptions of standard DID discussed above. To investigate the robustness of our results if some of these assumptions are relaxed and to assess the distributional effects of the reform, we now discuss the results obtained using the CIC model.

5.2. The CIC approach

In our CIC model, we linearly control for the set of covariates included in our benchmark specification.[34] Effectively, we first regress the math scores on covariates using ordinary least squares. Then, using the residuals, we estimate the effects of the reform using the CIC approach. Since DID with covariates and MDID mainly lead to statistically equivalent results, we assume that controlling for X linearly is not crucial for the results.[35]

In Table 7, four estimators are presented: (1) the DID model, (2) the discrete CIC model with conditional independence, (3) the discrete CIC model lower bound, and (4) the discrete CIC model upper bound.[36] The first column shows the mean effect. Columns 3, 5, 7 and 9 present the effects at the 25th, 50th, 75th and 90th percentile and thereby provide an overview of the distributional effect of the reform. Table 7 shows the empirical results for grade 2 (top panel), grades 3–4 (second panel), grades 5–6 (third panel), grades 7–8 (fourth panel) and grades 9–10 (bottom panel). All standard errors are bootstrapped to account for the clusters and stratifications of the NLSCY. The mean effects assuming conditional independence using CIC

[34] We decided not to control for the PPVT score for two reasons: (1) the results controlling for the PPVT scores are comparable and (2) the PPVT score is missing for about 20% of all our observations, which would greatly reduce our ability to detect distributional effects.

[35] Also, note that CIC estimates without covariates (not provided here, but available on request) are comparable to those with covariates.

[36] We modified the MATLAB program provided by Athey and Imbens to include the bootstrap weights provided by Statistics Canada to account for the sampling design of the NLSCY. We assume full responsibility for the computation of the estimates presented in this paper.

Table 7
Distributional effect of treatment on the treated.

Dependent variable: CAT/2 score	Mean	(Std. err.)	25th Perc.	(Std. err.)	50th Perc.	(Std. err.)	75th Perc.	(Std. err.)	90th Perc.	(Std. err.)
Grade 2										
Years 1996, 2000★										
DID	−20.25**	(7.98)	−11.88	(9.52)	−14.88	(12.98)	−40.19***	(11.19)	−25.10***	(5.81)
CIC dci	−17.04**	(7.51)	−14.00	(8.97)	−10.00	(12.60)	−28.00**	(12.44)	−13.00	(13.36)
CIC lower	−17.81**	(7.53)	−14.00	(9.08)	−10.00	(12.65)	−28.00**	(12.93)	−13.00	(14.12)
CIC upper	−16.48**	(7.50)	−14.00	(8.92)	−10.00	(12.60)	−25.00**	(12.34)	−13.00	(13.36)
Years 1996, 2006										
DID	−24.01***	(8.03)	−13.00	(7.92)	−25.14**	(12.61)	−46.54***	(11.29)	−40.14***	(6.73)
CIC dci	−21.22***	(8.00)	−8.00	(6.84)	−14.00	(14.84)	−37.00**	(14.74)	−31.00**	(13.25)
CIC lower	−22.02***	(8.00)	−9.00	(6.96)	−14.00	(14.81)	−37.00**	(14.71)	−39.00***	(13.31)
CIC upper	−20.58**	(7.99)	−8.00	(6.82)	−14.00	(14.89)	−37.00**	(14.85)	−28.00**	(13.55)
Years 1996, 2008										
DID	−10.20	(8.16)	−3.87	(8.12)	−10.93	(12.98)	−35.71***	(12.09)	−24.38*	(13.86)
CIC dci	−5.43	(7.92)	2.00	(7.31)	−1.00	(13.36)	−18.00	(13.48)	0.00	(18.14)
CIC lower	−6.23	(7.97)	−1.00	(7.49)	−1.00	(13.52)	−18.00	(13.51)	−5.00	(18.34)
CIC upper	−4.83	(7.89)	2.00	(7.27)	−1.00	(13.31)	−18.00	(13.63)	0.00	(18.49)
Grades 3–4										
Years 2000, 2002★										
DID	−19.42***	(4.33)	−15.49***	(5.54)	−18.49***	(4.27)	−16.35***	(4.17)	−23.87***	(7.07)
CIC dci	−15.17***	(4.18)	−16.00***	(5.91)	−17.00***	(4.77)	−11.00**	(4.76)	−15.00*	(8.10)
CIC lower	−15.59***	(4.19)	−16.00***	(6.04)	−18.00***	(4.90)	−11.00**	(4.91)	−15.00*	(8.19)
CIC upper	−14.71***	(4.18)	−16.00***	(5.92)	−17.00***	(4.82)	−11.00**	(4.74)	−15.00*	(8.08)
Years 2000, 2006										
DID	−16.38***	(4.76)	−23.58***	(5.83)	−17.21***	(5.29)	−8.58*	(4.77)	−10.25	(7.13)
CIC dci	−16.72***	(5.01)	−21.00***	(6.29)	−16.00**	(7.25)	−9.00	(5.84)	−11.00	(7.87)
CIC lower	−17.22***	(5.01)	−22.00***	(6.30)	−16.00**	(7.35)	−9.00	(5.98)	−11.00	(7.94)
CIC upper	−16.14***	(5.01)	−21.00***	(6.30)	−16.00**	(7.20)	−9.00	(5.78)	−9.00	(7.87)
Grades 5–6										
Years 2000, 2002										
DID	−13.44**	(5.97)	−15.49***	(4.94)	−9.37	(8.37)	−20.17***	(7.73)	−13.24	(8.51)
CIC dci	−9.47	(6.05)	−10.00*	(5.62)	0.00	(7.80)	−17.00	(10.63)	−16.00	(11.44)
CIC lower	−10.03*	(6.07)	−10.00*	(5.65)	0.00	(7.89)	−17.00	(10.68)	−16.00	(11.34)
CIC upper	−8.84	(6.04)	−10.00*	(5.65)	0.00	(7.78)	−17.00	(10.79)	−16.00	(11.55)
Years 2000, 2004★										
DID	−20.28***	(5.82)	−16.55***	(6.24)	−15.58**	(7.80)	−30.22***	(7.51)	−22.15**	(8.66)
CIC dci	−19.46***	(6.23)	−13.00*	(6.83)	−14.00	(9.03)	−35.00***	(10.32)	−25.00**	(11.80)
CIC lower	−20.09***	(6.23)	−14.00**	(6.98)	−14.00	(9.09)	−35.00***	(10.45)	−25.00**	(11.89)
CIC upper	−18.85***	(6.22)	−13.00*	(6.77)	−14.00	(9.02)	−32.00***	(10.24)	−24.00**	(11.77)
Grades 7–8										
Years 2000, 2006★										
DID	−33.47***	(7.42)	−23.38**	(11.43)	−39.20***	(9.28)	−30.52***	(10.01)	−25.19**	(9.91)
CIC dci	−29.82***	(7.36)	−28.00**	(11.24)	−39.00***	(10.28)	−22.00**	(10.74)	−6.00	(14.65)
CIC lower	−30.28***	(7.38)	−28.00**	(11.25)	−39.00***	(10.30)	−22.00**	(10.79)	−19.00	(14.73)
CIC upper	−29.24***	(7.34)	−24.00**	(11.26)	−39.00***	(10.30)	−21.00**	(10.70)	−6.00	(14.69)
Years 2002, 2006★										
DID	−22.73***	(6.75)	−26.05***	(6.43)	−25.03***	(7.25)	−10.19	(9.86)	−7.84	(7.35)
CIC dci	−23.74***	(7.34)	−25.00***	(7.58)	−24.00***	(8.75)	−12.00	(11.20)	−1.00	(12.78)
CIC lower	−24.35***	(7.36)	−26.00***	(7.55)	−26.00***	(8.85)	−12.00	(11.20)	−5.00	(13.13)
CIC upper	−23.13***	(7.33)	−25.00***	(7.61)	−23.00***	(8.75)	−12.00	(11.26)	−1.00	(12.68)
Years 2004, 2006★										
DID	−33.68***	(6.63)	−26.11***	(8.58)	−31.13***	(9.34)	−28.11***	(9.75)	−27.19***	(8.79)
CIC dci	−34.47***	(7.35)	−34.00***	(8.81)	−35.00***	(10.26)	−26.00**	(13.25)	−32.00*	(17.12)
CIC lower	−35.01***	(7.36)	−35.00***	(8.79)	−35.00***	(10.30)	−26.00*	(13.58)	−32.00*	(17.19)
CIC upper	−33.94***	(7.35)	−33.00***	(8.85)	−35.00***	(10.27)	−26.00**	(13.20)	−32.00*	(17.10)

Table 7 (*Continued*)

Dependent variable: CAT/2 score	Mean	(Std. err.)	25th Perc.	(Std. err.)	50th Perc.	(Std. err.)	75th Perc.	(Std. err.)	90th Perc.	(Std. err.)
Grades 9–10										
Years 2002 2008★										
DID	−44.13***	(10.72)	−49.95**	(21.96)	−40.50***	(13.44)	−38.42***	(13.22)	−21.65*	(12.51)
CIC dci	−43.51***	(11.11)	−66.00***	(23.37)	−46.00***	(16.78)	−47.00***	(15.85)	−24.00	(17.09)
CIC lower	−43.91***	(11.11)	−66.00***	(23.38)	−48.00***	(16.88)	−47.00***	(15.84)	−24.00	(17.11)
CIC upper	−43.13***	(11.10)	−65.00***	(23.46)	−46.00***	(16.78)	−47.00***	(15.86)	−24.00	(17.13)
Years 2004, 2008★										
DID	−28.12***	(10.54)	−33.69**	(14.65)	−29.69**	(11.73)	−25.68**	(12.07)	−16.13	(16.22)
CIC dci	−26.87***	(10.29)	−37.00***	(14.14)	−24.00	(15.12)	−27.00*	(14.22)	−16.00	(15.67)
CIC lower	−27.45***	(10.28)	−37.00***	(14.17)	−24.00	(15.15)	−27.00*	(14.14)	−17.00	(15.56)
CIC upper	−26.37**	(10.30)	−36.00**	(14.18)	−24.00	(15.23)	−27.00*	(14.40)	−16.00	(15.85)
Years 2006, 2008★										
DID	−33.56***	(11.12)	−28.05**	(13.32)	−28.47**	(12.55)	−24.05*	(13.65)	−31.96	(24.89)
CIC dci	−33.57***	(11.27)	−37.00***	(14.17)	−40.00***	(15.02)	−36.00**	(15.97)	−42.00*	(23.07)
CIC lower	−34.13***	(11.27)	−38.00***	(14.22)	−41.00***	(14.86)	−36.00**	(16.08)	−42.00*	(23.23)
CIC upper	−32.95***	(11.27)	−37.00***	(14.13)	−40.00***	(15.14)	−34.00**	(15.98)	−42.00*	(23.07)

Note: Shows the estimated distributional effect of the treatment on the treated on grade 2 students (top panel), grades 3–4 students (middle panel) and grade 5–6 students (bottom panel). Refer to Table 7's note on the next page for further details. *Note*: Shows the estimated distributional effect of the treatment on the treated on grade 7–8 students (top panel) and grade 9–10 students (bottom panel). The dependent variable is the CAT/2 score. DID refers to the difference-in-difference estimator in levels. CIC dci refers to the discrete changes-in-changes estimator with conditional independence, while CIC lower and upper refer respectively to the lower and upper bounds of the estimator. The first column shows the mean effect, while columns 3, 5, 7 and 9 show the distributional effect at the 25th, 50th, 75th and 90th percentiles. All specifications control for the set of covariates listed in the footnote of Table 5. To denote our main longitudinal cohort of treated students, we use "★" beside the year. Standard errors are bootstrapped to account for the clustering and stratifications of the NLSCY. Coefficient significance is denoted using asterisks: *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

are generally comparable to the DID estimates and the estimated bounds are fairly tight. Before we discuss the distributional effects, a few points on mean effects using CIC versus DID are worth mentioning.

5.2.1. Mean effects

First, estimated impacts on our main longitudinal cohort (students entering grade 1 in 1999 and 2000, denoted using "★" besides year) again suggest that the effect of the reform is increasing with exposure (except from grade 2 to grades 3–4). In grade 2, the mean effect is 17.0. It increases from 15.2 in grades 3–4, to 19.5 in grades 5–6, to 23.7–34.5 in grades 7–8, to 26.9–43.5 in grades 9–10.[37]

Second, in grades 5–6, for year 2000 versus 2002, the CIC estimate is small and negative, but not statistically different from zero (while the DID estimator is significant). As mentioned above, only grade 5 students were exposed to the reform and for only one year (grade 4, 2001). Comparison of the estimated impact of treatment for years 2000, 2002 versus 2000, 2004 for students in grades 5–6 also support the idea that longer exposure results in a higher impact. Students in grades 5–6 in 2004 have been exposed to the reform 5 years. The estimated impact on these students is negative and significant (on the order of 33.7% of a std. dev.). A similar pattern can be observed when comparing the impact of the reform on grade 2 students in 2000 (exposed 1 year) with those of 2006 (exposed 2 years).

Third, age at first exposure may be important. Comparing grades 5–6 students spending 0–1 year in the reform (years 2000, 2002) with those in grade 2 (years 1996, 2000), it appears that the reform had a significant negative impact on grade 2 students, but no impact on grade 5–6 students. Age at first exposure appears important, with younger children being more affected than older children. This finding needs to be interpreted with caution, as estimated effects on grade 2 students rely on observations with higher non response and only half of the students in grades 5–6 were treated.

Fourth, both CIC and DID support the idea that long term effects may differ from short term effects. We find that grade 2 students, 8 years after the implementation of the reform, no longer seem to experience a significant negative effect (the CIC and DID estimators for years 1998 versus 2008 is small and not different from zero).[38] The reform being ambitious, it is possible that it took a fair number of years for teachers to develop the necessary skills to fully deploy all aspects of the reform. It may also be the case that, observing the decline in students' academic performance, teachers informally decided to reintroduce some of their pre-reform teaching approaches, and set aside in part or in totality the reform approach. The NLSCY does not provide information on the actual teaching approaches, therefore we are unable to identify which of these two explanations is dominant. In any case, this finding implies that at best the provincial reform had no long run effects on the development of procedural mathematics skills. This conclusion is derived from one set of grade 2 students at one point in time and although math achievement is an important predictor of socioeconomic success, it is not the only one.

[37] Table A.2 in the Appendix presents the CIC estimates for grade 2 students using the 1998 observations. Results are generally smaller but remain comparable to those using 1996 observations.

[38] This can also be observed from Fig. A.1 in the Appendix.

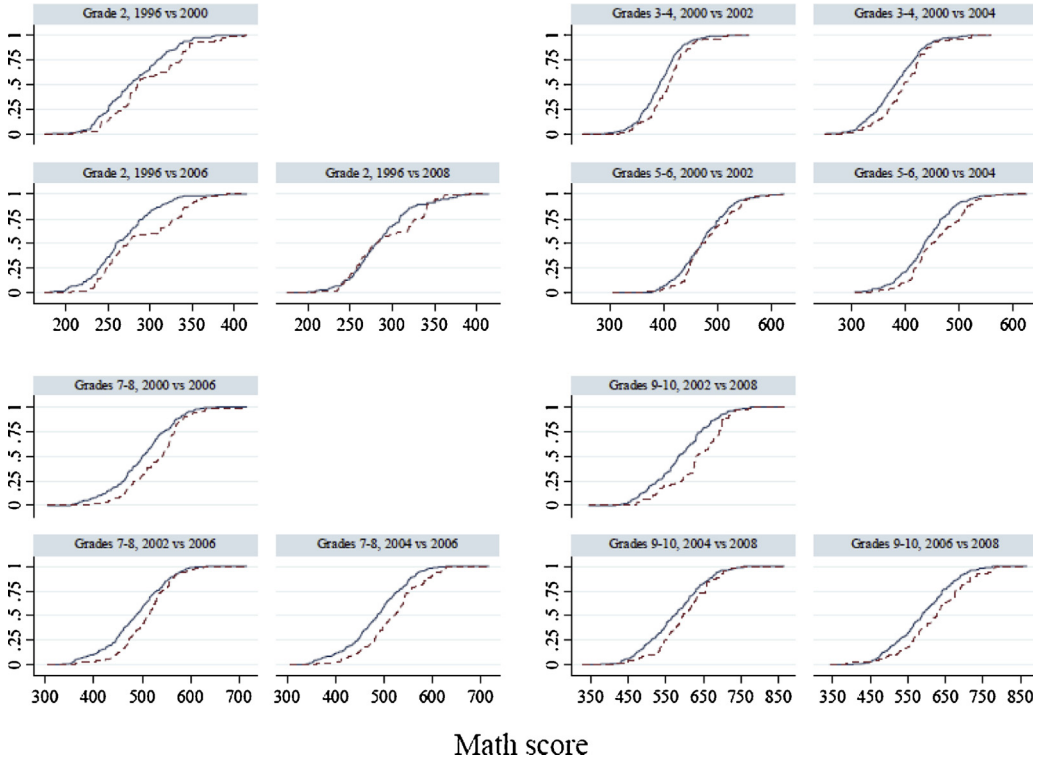


Fig. 2. Observed and counterfactual cumulative test score distribution function of Québec students post reform. Note: Shows the CAT/2 score cumulative distribution function of Québec students post reform. The solid line represent the observed distribution. The dashed line represent the counterfactual distribution estimated using the CIC approach. This figure represents graphically the results presented in Table 7: grade 2 students (upper left), grades 3–4 and 5–6 (upper right), grades 7–8 (bottom left), and grades 9–10 (bottom right).

5.2.2. Distributional effects

Looking across the entire math score distribution, we find that the results discussed above hold true for both lower performing students, and middle to top performing students.

In grade 2, only students in the 75th percentile appear to be significantly impacted by the reform. However as we move from grade 2 to grades 9–10, the effect also becomes significant for lower and average performing students. In grades 9–10, the magnitude of the coefficients is the largest for students in the 25th percentile, and slowly decreases as one moves toward the upper tail of the distribution. Looking at the top of the distribution (90th percentile), we also find negative effects across all grades, but the estimates are generally not significant.[39] It is possible that the reform did not harm top performers. It is also possible that the reform did impact top performers, but that the number of observations at this mass point is too small to obtain precise estimates. Fig. 2 shows the observed and counterfactual math score cumulative distribution function for Québec students post reform. This figure clearly shows that the effect is consistently negative across the entire distribution, and more so at the bottom of the distribution.

In sum, CIC suggests that the effects were negative on average and across the distribution. Lower performing students were impacted more severely, and the effects grew larger as students progressed from primary to secondary school. These large negative effects are worrying, and suggest that the reform may have harmed those most in need. Long term effects may be neutral. Further research is needed to fully understand the long term effects of the reform on a larger diversity of skills to get a better picture of the net benefit (or loss).

5.2.3. How material are these effects?

When discussing the CAT/2 test, we documented the influence of the test on later outcomes. We found that the CAT/2 score in grades 5–6 is associated with a marginal effect of −0.07% on the probability of dropping out of high school, a marginal effect of 0.12% on the probability of having a University degree, and a 25$ effect on total yeraly personal income at age 24–25. Going back to Table 7, we found that the mean effect on grades 5–6 scores (comparing years 2000 and 2004) was about −19.46. Assuming the marginal rate is constant, this decrease in the average math score would correspond to a 1.29% increase in the probability of dropping out of high school, a 2.39% decrease

[39] From the findings in Deke and Haimson (2006) discussed above, we were expecting top performers to not perform better as they were already at the high end of the distribution and further improving their skills was marginally more costly.

Table 8
Trend corrected distributional effect of treatment on the treated.

Dependent variable: CAT/2 score			25th		50th		75th		90th	
	Mean	(Std. err.)	Perc.	(Std. err.)	Perc.	(Std. err.)	Perc.	(Std. err.)	Perc.	(Std. err.)
Grades 5–6										
Years 2002, 2004, 2006										
DID-level	8.76		8.59		9.21		22.95		10.52	
CIC disc ci	3.13	10.68	8.00	11.92	−4.00	12.96	9.00	17.98	5.00	24.21
CIC disc lower	3.18	10.71	8.00	11.99	−5.00	13.01	10.00	18.00	2.00	24.63
CIC disc upper	3.20	10.65	9.00	11.97	−4.00	13.00	12.00	18.17	7.00	23.87
Grades 7–8										
Years 2002, 2004, 2006										
DID-level	−40.33		−18.88		−38.84		−58.32		−60.64	
CIC disc ci	−39.73***	14.00	−32.00*	17.34	−37.00**	18.47	−47.00**	20.27	−69.00**	29.87
CIC disc lower	−39.68***	14.03	−32.00*	17.40	−38.00**	18.47	−47.00**	20.39	−69.00**	29.88
CIC disc upper	−39.76***	13.97	−32.00*	17.40	−38.00**	18.57	−47.00**	20.17	−69.00**	29.88
Grades 9–10										
Years 2004, 2006, 2008										
DID-level	−48.08		−41.61		−34.22		−24.12		−41.54	
CIC disc ci	−49.33**	22.10	−57.00**	24.81	−53.00*	29.58	−45.00	32.88	−64.00	50.43
CIC disc lower	−49.30**	22.09	−56.00**	24.83	−52.00*	29.56	−45.00	32.86	−67.00	50.42
CIC disc upper	−49.42**	22.11	−58.00**	24.87	−54.00*	29.61	−45.00	32.97	−63.00	50.50

Note: Shows the estimated distributional effect of the treatment on the treated on grade 5–6 students (top panel), 7–8 students (middle panel) and grade 9–10 students (bottom panel). The dependent variable is the CAT/2 score. The structure of the table follows that of Table 5 and all specifications control for trends in addition to the set of covariates listed in the footnote of Table 5. Standard errors are bootstrapped to account for the clustering and stratifications of the NLSCY. Coefficient significance is denoted using asterisks: *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

in the probability of obtaining a University degree, and a 494$ decrease in total personal income at ages 24–25.[40]

One of the reform's objectives was to raise the proportion of students who successfully complete their high school education, which indirectly implies that the reform aimed at raising the achievement of lower performing students. Since mathematical ability (as measured by the CAT/2 test) is strongly related to school attainment and total personal income, the evidence presented above does not suggest that the reform achieved this objective.

5.2.4. Trends

The preceding sections have provided convincing evidence that the reform had negative effects on the math scores of students from the second to the tenth grade. However, some of this evidence was obtained with NLSCY cycles (1 and 3) with substantial non-response for the math scores. Also, it is feasible that some underlying pre-reform trends were already at work in both Québec and the Rest of Canada. The evidence based on incomplete data should not be dismissed as it clearly points in the same overall direction that the reform had negative effects on math scores. Estimated effects using imputed data are also in line with previous results.

To address the issue of possible pre-program trends in math by region and incomplete data in earlier cycles, we estimated the model using a triple difference approach with data from cycles 4 to 8 where response rates for math scores were very high, generally above 90% in Québec and 80% in the RofC. Because of these constraints, we cannot present estimates for children in grade 2, but only for grades 5 and 6, 7 and 8, and 9 and 10. The results appear in Table 8.

We observe that the estimated effects of the teaching reform remain strong, negative and significant for grades 7–10, but become small, positive and not significant for children in the fifth and sixth grade. But, again, for the latter group (grades 5–6), we compare students with a shorter exposure to the reform than the former groups (grades 7–10). In fact, it would be rather surprising that the reform would have an immediate large effect on skills in mathematics in grades 5–6.

Therefore, the overall evidence continues to demonstrate that the reform has reduced the skills of students in mathematics, in particular for grades in the latter stages of high school and certainly shows no positive effect on these same skills. Given the importance of mathematics in many fields and the amount of time and resources that were needed to implement this reform, these results should be disappointing for policy makers in the province of Québec and should concern policy makers considering reforms of the type implemented in Québec. Even though the effect of the reform on math scores is negative, the net effect of the reform depends on the overall impact of the reform on the development of a variety of skills. Although this is well beyond the scope of this paper, we provide some guidance on what the net effect might be by presenting the impact of the reform on other measures of achievement in Section 6.

6. Further evidence

6.1. Cognitive measures

The NLSCY is not the only source of information providing evidence on the reform. International assess-

[40] Murnane et al. (2006) show that basic skills in mathematics between the 1970s and 1980s have become increasingly important in predicting future wages. As mentioned above, they also show that basic mathematic skills predict wages better than other cognitive measures such as reading and vocabulary skills.

Table 9
Comparison of TIMSS performance across provinces.

	Mathematics achievement grade 4					Mathematics achievement grade 8				
Year	1995	1999	2003	2007	2011	1995	1999	2003	2007	2011
International	500	500	500	500	500	500	500	500	500	500
Québec	550	–	506*	519*	533*	556	566	543	528*	532*
Ontario	489	–	511	512	518	501	517	521	517	512
Alberta	523	–	–	505	507	–	–	–	–	505
British Columbia	–	–	–	505		–	522	–	509	–
	Coef.	Std. err.	*t*-stat	*N*		Coef.	Std. err.	*t*-stat	*N*	
DID estimate	−35.3	17.4	−2.03	11		−26.5	11.9	−2.23	12	
	Science achievement grade 4					Science achievement grade 8				
Year	1995	1999	2003	2007	2011	1995	1999	2003	2007	2011
International	500	500	500	500	500	500	500	500	500	500
Québec	529	–	500*	517*	516*	510	540	531	507*	520*
Ontario	516	–	540	536	528	496	518	533	536	521
Alberta	555	–	–	543	541	–	–	–	–	546
British Columbia	–	–	–	537		–	542	–	526	–
	Coef.	Std. err.	*t*-stat	*N*		Coef.	Std. err.	*t*-stat	*N*	
DID estimate	−22.1	16.8	−1.32	11		−22.0	9.1	−2.41	12	

Note: Mean scores were obtained from the Trends in International Mathematics and Science Study (TIMSS), years 1995, 1999, 2003, 2007 and 2011. Post reform years in Québec are denoted using an asterisk "*". DID estimates account for aggregated shocks at the provincial level (Wooldridge, 2003).

ments in which Canada participated can also be used for the analysis. Only a few of the ten provinces participated in the TIMSS.[41] This survey collects data on mathematical and science literacy and is administered to students in grades 4 and 8. The TIMSS mathematics test is broader than the CAT/2 test. In addition to procedural skills, the test also assesses ability in, for example, geometry, problem solving, reasoning and graphical representation.

The global scores by province are presented in Table 9. Grade 4 students are post-reform in years 2003, 2007 and 2011, and pre-reform in years 1995 and 1999, while grade 8 students are post-reform in year 2007 and 2011, and pre-reform otherwise. In mathematics and sciences, grade 4 students in Québec had a lower performance post reform (year 2003) compared to their performance in 1995. Scores in Ontario (Québec's neighboring province to the west) had in contrast increased over the same period. As of 2007 and 2011, the overall performance of Québec's 4th graders remained under its 1995 level, but had slightly increased compared to 2003. The performance in Ontario remained stable. Grade 8 students' performance shows a similar pattern when results from 2007 and 2011 are compared with results from all previous years: Québec's performance in both mathematics and sciences is trending downwards, while the performance in Ontario is increasing or stable.

Using the average scores by province, we estimated the DID estimator. This estimator is comparable to the estimator proposed by Donald and Lang (2007) and Wooldridge (2003). Estimated effects are large and negative in all cases. They are significant in mathematics in both grades, but only in grade 8 in science.[42] The standard errors obtained using this approach account for the possibility of aggregate random group-time specific effects. We argued before that this concern was unlikely to apply when analyzing mathematical test scores of children. As such, the estimated effects using TIMSS may be more significant than they appear here.

The ERES project (Evaluation of the Reform at the Secondary School Level) has recently produced two reports (August 2002, October 2012) comparing the math knowledge and the French proficiency of grade 11 students exposed to the reform to that of pre-reform ERES (2012a, 2012b). The math test uses 25 questions from exercises administered during the 2003 and 2006 PISA assessments. The test was administered to 1475 pre-reform students and 3189 post-reform students from 208 representative schools of Québec. In sum, students in the reform group scored slightly lower[43] on average, with a larger difference in geometry and algebra. As for the French proficiency test (coherent argumentation, vocabulary, sentence construction and punctuation, and spelling), they do not find any significant differences, but almost 30% of grade 11 students in the two groups (control 1180 and reform 1315) did not complete the assessment.

[41] Although all provinces in Canada have participated in the PISA since 2000, for the purpose of this analysis average results presented in official documents may not be easily compared. The only year post reform is 2009, which is also the year in which the response rate for the province of Québec (71%) was well below the international satisfactory threshold of 80% set by PISA (Knighton, Brochu, & Gluszynski, 2010, Table A.2). A non-response bias analysis conducted by Statistics Canada showed that students in less favourable socioeconomic environments were less likely to participate in PISA and that these students had a statistically lower performance on the provincial reading test.

[42] The same exercise using PISA data (with a 71% response rate post reform) leads to insignificant results in Reading, Mathemathics, and Science.

[43] They find a statistically significant difference of 1.5 percentage point.

Overall, the evidence from TIMSS and PISA suggests a worsening of Québec's students performance post reform in mathematics and at best a stand still in science and French. These results are in line with the more detailed results estimated using the NLSCY.

6.2. Behavioral measures

The focus of our paper was on mathematical skills. However, as mentioned above, the reform aimed not only at better preparing students from an academic stand point but also at developing active competencies such as communication skills, cooperation with others, etc. The NLSCY provides a number behavioral indicators. These outcomes are more subjective as they are the result of questions to the parent or the student.

Nonetheless, we estimated the impact of the reform on all behavioral outcomes available in the NLSCY (13 in total). The results not reported here can be viewed in our Web Appendix (Table A.3 presents the summary statistics, and Table A.4 the estimated effects). We find that the vast majority of coefficients across all grades and outcomes suggest a negative impact of the reform on students' behavior. With a few exceptions, these effects are rarely significant. In grades 5–6 and 7–8, we find a significant worsening of the situation for the following measures: hyperactivity, anxiety, physical aggression, interpersonal competencies and emotional quotient. In grades 5–6 and 7–8, the strongest evidence that the policy had an impact on behavior is for hyperactivity and anxiety (more than 50% of a std. dev.). The effects are rather strong, positive (more hyperactivity and anxiety) and robust to sample and method. In grades 9–10, the estimated effects are significant only for prosocial behavior, physical aggression and property offense.

Our results are in line with those reported by ERES (2011b) which found no effect on social adjustment, personal and emotional adjustment and intrinsic motivation. They also found that post-reform students felt less well-adapted to secondary school, male students were found to have lower self-esteem, and at risk students[44] were less engaged in school work. We therefore conclude that the reform did not improve the behavior of students measured using the self-reported NLSCY behavioral indicators.

7. Conclusion

We estimated the impact of the Québec school reform, based on a complete revision of teaching methods, on grade 2–10 students using math scores provided by the NLSCY. To our knowledge, no econometric evidence-based evaluation of the reform has been conducted to date. We find strong evidence of negative effects of the reform on the development of students' mathematical abilities. More specifically, using the changes-in-changes estimator, we show that the impact of the reform increases with exposure, and that it impacts negatively students at all points on the skills distribution. Results based on grade 2 students, suggest that long run effects may have been null. As such, the reform seems to have failed to meet one of its primary objectives (at least in the short to medium run). Students from the lower end of the distribution do not seem to be in a better position to successfully complete their schooling. Mathematical abilities are strongly related to school attainment and labor market outcomes, and for lower performing students they are at best equivalent post reform, but most likely lower.

The teaching approach dictated by the reform is based on socio-constructivism. According to Pinker (1997), proponents of this method believe that children must construct mathematical knowledge for themselves with the teacher only guiding the discussion on the topics and that repetitions and practice are seen as detrimental to learning. He argues that constructivism is not appropriate for mathematics. For him, "... without the practice that compiles a halting sequence of steps into a mental reflex, a learner will always be building mathematical structures out of the tiniest nuts and bolts". Certain skills for mathematics may be very difficult to "construct" at a young age and can possibly be better attained by old-fashioned practice and a more mechanical approach. Pinker suggests that the poor performance of the United States in mathematics could be linked to the teaching approach, which is mainly contextual with no teaching of mathematical concepts. The evidence presented in this paper supports this argument.

Mathematical skills are, however, not the only valuable skills that a student must develop in school. Although the debate is still ongoing on which skills should be developed in school, a consensus seems to have emerged on the importance of non-cognitive skills, or in other words, behavioral skills. Constructivism being heavily focused on communication and group interactions, it may be the case that the reform was better able to foster these skills. As pointed out by Deke and Haimson (2006), already high achieving students may have limited room to improve further in mathematics, but they may benefit from developing non-cognitive skills. The reform studied in this paper implemented a teaching approach that had a strong focus on non-cognitive skills such as communication, creativity and cooperation.[45] Using behavioral measures available in the NLSCY we did not find any evidence of the reform having a positive impact on behavioral outcomes, but we may still be missing part of the benefits (or losses) generated by the reform as these measures only capture part of the soft skills. It would be interesting to estimate the impact of the reform on the dropout rate, but comparable data across provinces on dropout rates for the first treated cohort is not yet available.

This paper highlights two important aspects of policy reforms. First, reforms of this magnitude have impacts

[44] Student risk status is determined by the parent and relates to aggressiveness, attention deficit and the level of prosociability.

[45] As mentioned above, Table A.1 in the Appendix provides the complete list of competencies and areas of learning.

beyond the scope of the original plan, and these effects should also be accounted for. While this may appear obvious to economists trained to think about general equilibrium effects, this is not always the case in other fields. In the present application, the reform's main target was to improve students' cross-curricular competencies, but it impacted at the same time the acquisition of specific skills in mathematics, and negatively so. Second, even if mastering the reform content may have taken some time, the impact on students during the transition period will always remain and should be taken into consideration when implementing large scaled reforms. While improving non-academic skills may be well placed and still require further analysis, we argue that the negative effects on mathematical skills remain worrying.

Acknowledgements

We gratefully acknowledge financial support from the Fonds québécois de la recherche sur la société et la culture (grant 2011-SE-137016). We thank the Québec Inter-University Centre for Social Statistics (QICSS) for great support throughout this project. The analysis is based on Statistics Canada's National Longitudinal Survey of Children and Youth (NLSCY) restricted-access Micro Data Files available at the QICSS. All computations on these micro-data were prepared by the authors who assume responsibility for the use and interpretation of these data. We thank for helpful comments participants of seminars at McGill, UQAM (Department of Mathematics), and at the Meetings of the CEA (University of Calgary), SCSE (University of Sherbrooke) Population Conference (University of Stockholm). The comments of two anonymous referees and the editor were extremely valuable in improving this work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.econedurev.2014.03.004.

References

AIR (2005). *Report on elementary school comprehensive school reform models. Comprehensive school reform quality center*. Washington, D.C. American Institutes for Research.

Angrist, J., & Krueger, A. (1999). Empirical strategies in labor economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (pp. 1277–1366). Elsevier: North Holland.

Angrist, J., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics, 114*, 533–575.

Angrist, J., & Lavy, V. (2002). New evidence on classroom computers and student learning. *Economic Journal, 112*, 735–765.

Athey, S., & Imbens, G. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica, 74*(2), 431–497.

Bailey, M., & Borooah, V. (2010). What enhances mathematical ability? A cross-country analysis based on test scores of 15-year-olds. *Applied Economics, 42*, 3723–3733.

Baker, M., Gruber, J., & Milligan, K. (2008). Universal child care, maternal labor supply, and family well-being. *Journal of Political Economy, 116*(4), 709–745.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics, 119*(1), 249–275.

Besley, T., & Case, A. (2000). Unnatural experiments? Estimating the incidence of endogenous policies. *Economic Journal, 110*(467), F672–F694.

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources, 44*(3), 565–640.

Borman, G., Hewes, G., Overman, T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(1), 125–230.

Bulle, N. (2011). Comparing OECD educational models through the prism of PISA. *Comparative Education, 47*(4), 503–521.

Bowlby, J., & McMullan, K. (2002). *At a Crossroads: First Results for the 18 to 20-Year-old Cohort of the Youth in Transition Survey*. Statistics Canada Catalogue no. 81-591-XPE.

Chetty, R., Friedman, J., & Rockoff, J. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. NBER Working Paper # 17699*.

Crawford, D., & Snider, V. (2000). Effective mathematics instruction the importance of curriculum. *Education and Treatment of Children, 23*(2), 122–142.

Currie, J., & Thomas, D. (2001). Early test scores, socioeconomic status and future outcomes. *Research in Labor Economics, 20*, 103–132.

Deke, J., & Haimson, J. (2006). *Valuing student competencies: Which ones predict postsecondary educational attainment and earnings, and for whom?* Princeton, NJ: Mathematical Policy Research (submitted to: Corporation for the Advancement of Policy Evaluation).

Donald, S., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *Review of Economics and Statistics, 89*, 221–233.

ERES (2012a). The math knowledge of education reform students. *Progress Report on Math, 4*(1), 1–2.

ERES (2012b). The written French proficiency of education reform students. *Progress Report on French Proficiency, 4*(2), 1–2.

ERES (2011a). Teaching practices, classroom climate and usefulness of classes as perceived by students. *Progress Report to Chart Knowledge Acquisition, 3*(1), 1–2.

ERES (2011b). A few indicators of students' socio-motivational profile. *Progress Report to Chart Knowledge Acquisition, 3*(2), 1–2.

Haeck, C., Lefebvre, P., & Merrigan, P. (2013). *Canadian Evidence on Ten Years of Universal Preschool Policies: The Good and the Bad*. CIRPÉE – Working Paper 2013–17: http://www.cirpee.org

Hanushek, E. (2003). The failure of input-based schooling policies. *Economic Journal, 113*, F64–F98.

Hanushek, E., & Rivkin, S. (2010). The quality and distribution of teachers under the no child left behind act. *Journal of Economic Perspectives, 24*(3), 133–150.

Hanushek, E., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature, 46*(3), 607–668.

Heckman, J., Ichimura, H., & Todd, T. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies, 65*(2), 261–294.

Heckman, J., Ichimura, H., & Todd, T. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies, 64*(2), 605–654.

Heckman, J., & Rubinstein, J. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review, 91*(2), 145–149.

Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411–482.

Hoxby, C. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics, 115*(4), 1239–1285.

Ingram, B., & Neumann, G. (2006). The returns to skill. *Journal of Labor Economics, 13*, 35–59.

Jackson, E., & Page, M. (2013). Estimating the distributional effects of education reforms: A look at project STAR. *Economics of Education Review, 32*, 92–103.

Knighton, T., Brochu, P., & Gluszynski, T. (2010). *Measuring Up: Canadian Results of the OECD PISA Study. The Performance of Canada's Youth in Reading, Mathematics and Science 2009: First Results for Canadians Aged 15. Statistics Canada, Catalogue no. 81-590-XPE – No. 4*.

Le, V.-N., Stecher, B., Lockwood, J., Hamilton, L., Robyn, A., Williams, V., Ryan, G., Kerr, K., Martinez, J., & Klein, S. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement*. Santa Monica, CA: RAND Corporation, http://www.rand.org/pubs/monographs/MG480

Machin, S., & Mcnally, S. (2008). The Literacy Hour. *Journal of Public Economics, 92*, 1441–1462.

MELS (1999). *The Education Reform What It's All About. Gouvernement du Québec, Ministère de l'Éducation, du Loisir et du Sport*. http://www.mels.gouv.qc.ca/reforme/mieux_enfants/dépcoul_a.pdf.

MELS (2001). *Québec Education Program: Preschool Education, Elementary Education. Gouvernement du Québec, Ministère de l'Éducation, du Loisir et du Sport*. http://www.mels.gouv.qc.ca/GR-PUB/m_englis.htm

MELS (2003). *Québec Education Program: Secondary Cycle One. Gouvernement du Québec, Ministère de l'Éducation, du Loisir et du Sport*. http://www.mels.gouv.qc.ca/GR-PUB/m_englis.htm

MELS (2005). *Education in Québec. An Overview. Québec: Ministère de l'Éducation*. http://www.mels.gouv.qc.ca/scolaire/educqc/pdf/educqceng.pdf

MELS (2007). *Québec Education Program: Secondary Cycle Two. Gouvernement du Québec, Ministère de l'Éducation, du Loisir et du Sport*. http://www.mels.gouv.qc.ca/GR-PUB/m_englis.htm

Mourshed, M., Chijioke, C., & Barker, M. (2010). *How the world's most improved school system keep getting better*. McKinsey & Company.

Murnane, R., Willett, J., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics, 77*(2), 251–266.

Murnane, R., Willett, J., Duhaldeborde, Y., & Tyler, J. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management, 19*(4), 547–568.

Neal, D., & Johnson, W. (1996). The role of pre-market factors in black-white differences. *Journal of Political Economy, 104*(5), 869–895.

Pinker, S. (1997). *How the mind works*. New York: WW Norton & Compagny.

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rose, H., & Betts, J. (2004). The effect of high school courses on earnings. *Review of Economics and Statistics, 86*(2), 497–513.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rouse, C., & Krueger, A. (2004). Putting computerized instruction to the test: A randomized evaluation of a "scientifically based" reading program. *Economics of Education Review, 23*(4), 323–338.

Ryan, C. (2013). What is behind the decline in student achievement in Australia? *Economics of Education Review, 37*, 226–239.

Schafer, J. (1997). *Analysis of incomplete multivariates data*. London: Chapmand and Hall/CRC Press.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. Falmer Press: London.

Wooldridge, J. (2003). Cluster-sample methods in applied econometrics. *American Economic Review, 93*(2), 133–138.