KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Department of Economics

All students left behind: an ambitious provincial school reform in Canada, but poor math achievements from grade 2 to 10
by

Catherine HAECK
Pierre LEFEBVRE
Philip MERRIGAN

Econometrics

Center for Economic Studies
Discussions Paper Series (DPS) 11.28
http://www.econ.kuleuven.be/ces/discussionpapers/default.htm

October 2011

DISCUSSION PAPER

All students left behind: An ambitious provincial school reform in Canada, but poor math achievements from grade 2 to 10.

Catherine Haeck, Pierre Lefebvre and Philip Merrigan*

October 2011

Abstract

We investigate the impact of an ambitious provincial school reform in Canada on students' mathematical achievements. This reform provides advantages for the purpose of evaluation and cuts across some of the methodological difficulties of previous research. First, the reform was implemented in every school across the province in both primary and secondary schools. Second, we can differentiate impacts according to the number of years students are affected by the reform. Third, our data set provides a longer observation period than typically encountered in the literature. We find negative effects on students' mathematical achievements at all points of the skills distribution.

* *Catherine Haeck:* Katholieke Universiteit Leuven. Email: Catherine.Haeck@econ.kuleuven.be. *Pierre Lefebvre:* Université du Québec à Montréal. Email: Pierre.Lefebvre@uqam.ca. *Philip Merrigan:* Université du Québec à Montréal. Email: Philip.Merrigan@uqam.ca. We gratefully acknowledge financial support from the Fonds québécois de la recherche sur la société et la culture and from the Belgian Federal Science Policy Office (Interuniversity Attraction Poles P5/26). 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.

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

The expansion of international assessments of students in school and the growing number of countries participating in these surveys[2] have provided detailed pictures of academic achievements with comparative performance measures across countries. Four important points stand out from the empirical results. First, there are significant international differences in overall tests scores, even among high-income countries. Second, there are important disparities in the results between students within the same country. Third, countries performing strongly generally display the smallest disparities in the results (Knighton, Brochu, and Gluszynski 2010; Gonzales et al., 2008). Fourth, social background is a strong determinant of student achievement in a number of countries (Fuchs and Woessmann, 2007; Bussière, Catwright and Knighton, 2004; Fleischman et al. 2010). Improving the performance of low-skill students can help reduce the overall disparity in scores between well-off and deprived students, and increase the overall performance of the country.

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

[2] For examples of repeated international surveys see: the OECD's Program for International Student Assessment (PISA 2000, 2003, 2006, and 2009) in the domains of reading, mathematical and science literacy administered to 15-year-olds; the International Association for the Evaluation of Educational Achievement who has conducted five international Trends in Mathematics and Science Studies (TIMSS 1995, 2003, 2007, 2011) in the domain of mathematical and science literacy administered to students in grade 2 and grade 8; and the three Progress in International Reading Literacy Study (PIRLS, 2001, 2006, 2011) administered to 10-year-olds. In the 2009 PISA survey, 65 countries/regions (up from 43 in 2000) all over the world participated in the assessment, including Hong Kong and Shanghai (China).

A consensus seems to have emerged from these surveys (and the research on education production function) suggesting that a sizeable proportion of young people around age 15 in many countries do not appear to possess all of the skills required to meet the challenges of today's knowledge societies. As a result, in recent years, public pressure to upgrade educational standards and improve academic achievements has been steadily increasing in many countries. Education authorities have responded in a number of ways.[3]

In the United States, a number of reforms were implemented, with comprehensive school reform (CSR)[4] and charter schools[5] leading the way. The potential of CSR to improve children's performance remains unclear, with most studies showing only modest effects - or sometimes no effect - on student achievement (Vernez et al. 2006; Orland, Hoffman and Vaugh, 2010; Borman et al., 2003). Findings from recent studies on charter schools, (reviewed in Gleason et al., 2010) based on non-experimental methods, have also been mixed. Research on CSR models and charter schools are limited in a number of ways. First, research focused on implementation finds that schools receiving grants failed to implement the full model (Vernez et al. 2004) and were not more likely to have implemented the reform compared to other schools five years after receiving the grant (Orland et al., 2010). Yet, a majority of studies simply assume that a grant-receiving school implements the model. Second, a majority of schools involved in these reforms exhibit higher than average poverty rates, such that results found using these reforms may not be transferable to lower poverty rate settings. Third, the variety of reforms implemented, the choice of students treated, the different possible financing mechanisms and the various geographical locations create considerable heterogeneity that is empirically difficult to address in order to provide convincing evidence on the most promising reforms.

[3] For some recent European reforms in response to the PISA surveys see Grek (2009) and Ertl (2006).

[4] CSR models typically impact all aspects of school operations. See Borman et al. (2006) for a complete list of the components used by the U.S. Department of Education to define CRS. We further detail CRS in Section 2.

[5] The US federal government supports the charter school movement. Charter schools which operate under a contract (charter) with a government agency that subsidize them are provided a degree of autonomy from local school boards, and have fewer regulations with additional accountability requirements.

In this paper, we estimate the impact of Québec's (the second most populated province of Canada) ambitious and universal school reform implemented in the early 2000's on children's mathematical ability throughout primary (K-6) and secondary (7-11) school. At the time of the reform, Québec was among the top performing countries in international assessments, but it was still subject to severe criticism at home due to its alarmingly large high school dropout rate, especially among male students.[6] To ensure success for all, the province decided to implement an ambitious reform introducing a new curriculum 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, and 2007) relied on a competency-based approach. It moved teaching away from the traditional/academic approaches of memorization, drills and activity books, to a much more comprehensive approach focused on learning in a contextual setting in which children are expected to find the answers for themselves. The Québec experiment/reform provides some advantages for the purpose of evaluation and cuts across some of the methodological difficulties mentioned above. First, Québec's Department of Education implemented the reform and all schools (public and private)[7] were forced to apply the new education program. Second, the reform's curriculum content was supported by a number of countries. Evidence from Bulle (2011) suggests that most OECD countries are moving away (or have long moved away) from the traditional (more academic) teaching approach. Yet it remains unclear whether the traditional approach is preferable or not to the contextual approach focused on the development of competencies. The Québec school reform can provide direct empirical evidence on this question.

The Québec school reform was implemented in steps, starting in September 2000 for grades 1 and 2, ending in September 2008 for grade 11 (i.e. the last grade of high school in Québec). Teaching in the Rest of Canada (RofC) continued to be delivered in the same

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

[7] Private schools are highly-subsidized in Québec and must adhere to the government's requirements to receive the subsidies.

way throughout the period. Therefore, it is possible to estimate the effect of the reform using students from the other provinces, who are in the same grade as Québec's students, as controls. Furthermore, comparing over time "treated" younger students (newly exposed to the reform) and older students (exposed for many years to the reform) to a comparable "control" group of students allows us to assess both the impact of the reform on mathematical ability throughout primary and secondary school and the impact of the reform for different lengths of exposure to treatment.

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 abilities 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 and Krueger, 1999; Blundell and 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. The CIC framework allows us to estimate the impact of the reform at different points on the skills distribution. More specifically, we can investigate whether or not the reform had a positive impact on the least performing students. Research by Deke and Haimson (2006) suggests that, if the reform was effective, we should find strong gains in mathematics on low achievers and but possibly no gains on high achievers. They find that the benefit of incremental gains in a competency does depend on the mix of skills each student possesses such that it is more effective for students to improve in areas where they are weak than to focus on further developing areas where they are well above average.[8] More specifically, they show that the benefit of improving math test scores appears much greater for students who are weak in math. Using CIC, we also estimate the impact on high achievers.

[8]The authors examine how several indicators of academic and non-academic competencies, specifically, indicators of mathematics skills, work habits, leadership skills, teamwork and other sports-related skills, and attitudes toward whether luck or effort determines success in life ("locus of control"), are related to postsecondary education (of enrolling and completing a program) and labour market outcomes.

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 and that the effects were larger the longer the exposure to the reform.

The outline of the paper is as follows. Section 2 provides a brief review of recent research on school reforms. Section 3 highlights the distinctive features of the education system in Québec compared to the other Canadian provinces, and describes the school reform implemented since 2000. Section 4 exposes the econometric methodology used to identify the causal effect (treatment on the treated effect) of the school reform on mathematics achievement. Section 5 describes the data set used and presents descriptive statistics of the key variables. The estimated effects of the reform are presented in Section 6. The last section discusses the results and conclusions.

1 Review of research on school reforms

In the United States, improving children's academic achievement in public schools was addressed through the prism of race/ethnicity score gaps. Thus, a major objective of government policy has been to deal with the black–white achievement gap by improving the quality of elementary and secondary level education for disadvantaged students with such actions as desegregation and school finance redistribution. However the persistent gaps raise questions about the effectiveness of these school interventions (Hanushek and Rivkin, 2009).

CSR has 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. Though not currently encouraged systematically by federal policy, CSR has a history of federal support (AIR 2005; Berends, Bodilly, and Kirby 2002). Proven approaches (i.e. research-based CSR) were supported by federal grants through the CSR Program (CSRP). Funding is generally given to schools in high-poverty areas and with lower student performance first. Borman et al. (2003) state that the average poverty rate of schools receiving federal funds was around 70% at the time of their study.

A number of studies suggests that CSR may have positive effects on student outcomes, and that effective implementation of a research-based CSR model is the key factor for success (Bifulco, Duncombe and Yinger, 2005; AIR, 2005; Vernez et al. 2006; Klugh and Borman, 2006; Aladjem et al., 2006). A meta-analysis of over 230 evaluations of individual CSR models (half of which were conducted by the model developers themselves) on 29 leading CSR models across the country found that their overall effects are positive, but small[9] (Borman et al., 2003). Vernez et al. (2006) focus on implementation and find that most CSR schools failed to implement the full model, which may explain why most of the research to date has found at best small effects. Furthermore, looking at the long term effect of CSR grants, Orland et al. (2010) compare grant recipients with other schools and find that grant recipients were not more likely to implement the CSR legislative specifications five years after receiving the grant. The authors also show that, five years later, these schools did not exhibit greater improvement in mathematics and reading achievement.

Of the few experimental studies on CSR, one clearly relates to the Québec reform curricula. Crawford and Snider (2000) study two curricula: one explicitly teaches mathematical concepts and focuses on the mastery of mathematical concepts through drill and repetition,

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

the other uses a more implicit approach in which the teacher sets up a situation and the students have to learn and discover concepts through reasoning and discussions, and it 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.

General conclusions about CSR programs are inherently difficult to provide given the variety of programs available and the failure to measure implementation. Furthermore, given the particularities of the CSR grant allocation process, general results may not be directly transferable to lower poverty rate settings or even to similar settings with lower commitment to improve the outcomes of students.[10]

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 nonsignificant or at best weakly positively significant. Although extensive work was deployed to gather classroom data revealing the implemented teaching approach, the data collected and the contextual setting pose a number of limitations. First, since students were not randomly assigned to teachers, most students experienced a mix of teaching approaches

[10]The charter school movement is another approach currently in place in the United States whose aim is to increase school effictiveness. This reform relates more to the organisation of the school system (centralized versus decentralized) than to the curriculum content and application. These schools serve mainly under-privileged students from low-income family from inner-cities. The evidence on the effects of charter schools is mixed. Studies covering a wide span of states and/or districts, found no impacts in reading and mathematics. Experimental studies focusing on large urban areas serving large populations of disadvantaged and racial/ethnic minority group students found positive impacts (Gleason et al., 2010; Hoxby, Murarka, and Kang, 2009; Abdulkadiroglu et al., 2009)

during the three year observation period. Second, teachers self-selected into the implementation of the reform-oriented approach, such that the estimated impact cannot reveal the impact of the reform if applied to all teachers. Third, the teaching approach was mainly self-reported by teachers, and somewhat conflicting evidence was found through observation of a few of the sampled teachers. Fourth, teachers admitted to being influenced by the testing environment. Given the push forward for similar teaching approaches in the United States, combined with limited empirical evidence concerning the effectiveness of these approaches in raising students' achievements, further research is required.

In sum, reforms in the United States in recent years have followed a number of different approaches, but targeted mainly disadvantaged children. Results to date remain mixed, but somewhat on the weakly positive side.

2 Québec's school system and curriculum 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 (with a K-11 rather than a K-12 system).

In all of Canada, children start school in kindergarten at age 5 in most cases, but sometimes as early as age 4 depending on the provincial regulation concerning entry age. Children then move on to primary school, where they complete six years of education from grade 1 to 6. Children then pursue their education in high school. In Québec, high school consists of five years of education, grades 7 to 11, while in the RofC children must complete six years of education, grades 7 to 12, to obtain their high school diploma. Grades 7 to 11 in the RofC are comparable to those in Québec.[11]

[11] Our data set, further detailed in Section 5, covers grades 2 to 10 students.

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 behaviours" (MELS, 2005).

In the classroom, what should have been different? 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 is mandatory in each and every school, funding is not tied to the implementation, and training packages and support is centralized in many ways.

The allocation of time per subject was also modified[13]. More time was spent on learn-

[12] A complete list of the competencies and areas of learning is provided in Table 9 in Appendix.

[13] The main areas and subjects of the curriculum are: 1. Languages (French or English as a teaching language, and French/English as a second language). 2. Mathematics, science, technology. 3. Arts education (art, music, drama or dance). 4. Physical education and health. 5. Moral education, or Catholic religious and moral instruction or Protestant moral and religious education.

ing the language of instruction (French or English) and mathematics, while less time was spent on all other subjects. More specifically, in high school some subjects were completely dropped (e.g. home economics), while others were integrated in the curriculum of other broader subjects (e.g. economics with citizenship education, human biology with science and technology).

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, drill and traditional lectures in which teachers provide the content to be learned appears to have been put aside.

2.2 The implementation

Figure 9 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 2002 - grades 5 and 6 (Elementary Cycle 3); September 2004 and 2005 - grades 7 and 8 respectively (Secondary Cycle 1); September 2006, 2007 and 2008 - grades 9, 10 and 11 respectively (Secondary Cycle 2). 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 the above timeline, and that parents were not able to self-select their children into or out of the reform, except by moving out of the province.

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 reformed curriculum despite the fact that it did not affect all levels of schooling at the time.

The NLSCY does not provide any information on the extent to which the reform was implemented in the school attended by the child. As the reform was mandatory, we can safely assume that at least part of the reform was enacted in each school.

2.3 International comparability

The curriculum content of the Québec reform resembles the reform-oriented teaching approach discussed above in Le et al. (2006) in the United States. In contrast to this reform, the Québec reform was mandated to each and every school across the province by the Department of Education, such that every school and teacher had to embrace the reform (at least to some level). Children were either treated or not, while in Le et al. children could be treated one year, not treated the next, and treated again the year after. Finally, our data set provides a much longer observation period than the three year period covered in Le et al. As a result, children can be observed longer, such that the observed number of years into treatment is larger. Also, prior research by 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).

As of 2006, the reform-oriented teaching 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. As such, findings related to Québec's school reform can contribute to debates across the border.

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 Costas Dias (2009) describe applications and give an 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, and Mullainathan, 2004; Donald and Lang, 2007; and Besley and Case, 2000). As a result, in addition to standard DID, the changes-in-changes (CIC) model developed by Athey and Imbens (2006) and the matching difference-in-differences estimator developed by Heckman, Ishimura and Todd (1997, 1998) are also used to estimate the impact of the reform.

The CIC model relaxes some of the assumptions of the standard DID.[14] Standard DID assumes outcomes are additive in time period, group and unobservable characteristics of the individual, while the CIC model is nonparametrically identified.[15] 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 unobservable

[14] Note that the standard DID is a special case of the CIC model.

[15] 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.

characteristics of individuals may differ across groups and the treatment effect may also differs 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 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. Using CIC, it is thus possible to evaluate the effect of a policy intervention not only in terms of the mean effect, but also in terms of the quantiles of the distribution.[16] This feature is particularly attractive in the present application, as knowing whether lower performing students benefited more or less, as compared to middle to top performing students, is of great policy interest.[17]

The CIC model relies on the assumption that the underlying production functions for treated individuals and non-treated individuals, mapping the relationship between the outcomes and the unobservable characteristics at a given point in time, do not vary across groups. As long as this holds true, CIC provides consistent estimates of the effect of treatment on both treated and untreated groups of 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. It is thus possible to estimate the effect of treatment at different points in the distribution.[18] As such, CIC permits policy evaluations in terms of 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 implemen-

[16] Quantile DID, which applies DID to each quantile as opposed to the mean, can also look into the distributional effects of the treatment. In line with standard DID, it assumes that the underlying distributions of unobservable characteristics of the individuals must be the same in all subgroups. As discussed above, the CIC model has the advantage of allowing for heterogenity across groups.

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

[18] See Athey and Imbens (2006) for the benefits of CIC over quantile DID.

tation 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:

$$\tau^{CIC} \equiv E\left[Y_{11}^{I}\right] - E\left[Y_{11}^{N}\right] = E\left[Y_{11}^{I}\right] - E\left[F_{Y,01}^{-1}\left(F_{Y,00}\left(Y_{10}\right)\right)\right], \tag{1}$$

where Y_{gt}^{I} is the outcomes of students receiving treatment in group g in time period t and Y_{gt}^{N} is the outcomes of students not receiving treatment in group g in time period t. In equation 1, Y_{11}^{I} is the outcomes of students receiving treatment in Québec after the implementation of the school reform and Y_{11}^{N} is the outcomes of students not receiving treatment in Québec after the implementation of the school reform. Y_{11}^{N} is not observed, but can be inferred using $E\left[F_{Y,01}^{-1}\left(F_{Y,00}\left(Y_{10}\right)\right)\right]$. $F_{Y,01}$ 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 outcomes of students in Québec prior to the reform. Standard errors are bootstrapped to account for the sampling design of the NLSCY. Individual characteristics (denoted X) need not be stable over time or across subpopulations, as long as the changing characteristics are observed.

Deke and Haimson (2006) show that some students are more likely to benefit from an improvement in academic competencies, and 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. The CIC also allows us to estimate the impact of the reform at different points in the skill distribution. This overall distribution is obtained using the three distributions of outcomes observed: treated before treatment, control before treatment and control after treatment. In our setting, each point on the distribution is inferred as follow. 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.

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 et al, 1997, 1998). This estimator also allows the possibility of selection into treatment. Linguistic differences and spatial distances between Québec and the RofC makes this possibility very unlikely.[20] However, the variation in response rates across waves (especially waves 2 and 3), may create a selection bias. With repeated-cross sections, the MDID estimator is (Blundell and Costa Dias, 2009):

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

where individual 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 individuals in S_{10} and S_{00}. The outcome variables are measured at time t_1 (after the reform) for individuals in S_{11} and S_{01}. Each individual j when compared to individual i is attributed a specific weight $\tilde{w}_{ijt}$ that depends on the matching technique used, and w_i stands

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

[20] As mentioned above, in the present application selection into treatment is extremely unlikely, since the only way to self-select out of (or into) treatment is to change the family's province of residence.

for sampling weights. The MDID estimator controls for X non-parametrically by ensuring that students in each group (control prior to treatment, control after treatment and treated prior to treatment) all share the treated group after treatment distribution for each of the characteristics contained in X.

Effectively, we first estimate a probit model in which the dependent variable equals one if the student lived 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 for local linear regression and kernel matching to account for the underlying matching procedure (not consistent for nearest neighbor). 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. We find that the estimated confidence intervals considerably overlap, such that linearly controlling for the students and family characteristics in our CIC approach is not overly restrictive.

4 Data set

The data set used for our empirical analysis is Statistics Canada's National Longitudinal Survey of Children and Youth (NLSCY), 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 behaviour of the child as well as data on the demographic situation 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 population of Canadian children aged 0 to 11 years old, first selected in wave 1 (1994-1995) of the survey.[21]

In wave 1, a sample of 22,831 children was selected. This sample constitutes the main longitudinal sample of the NLSCY. To reduce the response burden on families with several eligible children, the number of children selected per family was limited to two in wave 2 (1996-1997). As a result, some children were dropped from the original sample and 16,903 children remained in the longitudinal sample. The rule changed again to one child per household in wave 5 (2002-2003). Given the timeline of the reform, this change implies that a sibling fixed effects method cannot be used to evaluate the impact of the reform using the NLSCY. At the time the NLSCY survey was last conducted (wave 8, 2008-2009) longitudinal children were aged 14 to 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 to 6, or 6 to 11 year olds) to 2008-09 (grades 9 to 12, or 14 to 17 year olds). Later on, a new initiative providing additional observations on primary school students in grades 1 to 4 in academic year 2006-07, and in grades 1 and 2 in academic year 2008-09, was added to the main longitudinal survey.

In sum, the NLSCY provides three cohorts of children of primary and secondary school age: (1) students in grades 1 to 6 in academic year 1994-95 up to grades 9 to 12 in academic year 2008-09, (2) students in grades 1 to 4 in academic year 2006-07, and (3) students in grades 1 and 2 in academic year 2008-09.

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

4.1 Test scores

The NLSCY provides one measure of cognitive development for school age children: the CAT/2 mathematics test[22]. The CAT/2 test is a shorter version of the Mathematics Computation Test taken from the Canadian Achievement Tests, 2^{nd} edition. This test was developed by the Canadian Test Centre after careful consideration of the differences across the main school curricula across Canada. The CAT/2 is designed to measure basic skills in mathematics (addition, subtraction, multiplication, division on integers, etc.). The test is administered to students enrolled in grades 2 to 10, aged 7 to 15 years old. 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.[23] 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 2^{nd} graders may be as proficient in mathematics as some lower performing 3^{rd} 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 wave, and that students of the same age passed different tests. Because the difficulty level of the test for comparable students is different in wave 1 (compared to all other waves), we exclude observations from wave 1 from our analysis.[24]

[22] In waves 2 and 3, a reading test was also administered (in addition to the CAT/2 mathematics test). However, as of wave 4 the reading test was discontinued because of time constraints.

[23] 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.

[24] In wave 1 (1994-95), only 3 levels of the test were available: one for students in grades 2 and 3, one for grades 4 and 5, and one for grades 6 and 7. In this first wave of the NLSCY, a significant number of

In waves 2 and 3, the test was administered by the student's teacher. The student's parent had to sign a consent form, and the School Board and the teacher had to agree on taking the time to administer the tests. The response rates for these waves were uncharacteristically low: 74% in wave 2, and 54% in wave 3. From wave 4 onward, to avoid disrupting class activities at the end of the school year, the math test was administered at home by the interviewer rather than at school, and almost all eligible students (approximately 90 percent) responded. In most of our empirical work, we rely on test scores taken from wave 4 onwards, which covers pre- and post-reform students in most instances. However, for grade 2 students, wave 4 is already post-reform. Therefore we must use wave 2 and 3 observations to infer the impact of the reform on these students. Given the low response rate in wave 3, results using this wave are to be interpreted with caution.

In wave 3, students in grades 9 and 10 were observed for the first time. At the time, a single test was administered to these students. In wave 5, to better assess the development of students in grades 9 and 10, Statistics Canada decided to create separate tests for grades 9 and 10. Scores obtained in wave 5 and above cannot be compared with scores obtained in waves 3 and 4. As a result, we do not estimate the impact of the reform using test scores prior to wave 5 for grades 9 and 10.

4.2 Students observed and the incidence of the reform

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. 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. More generally, the entire school curriculum was designed by cycle. As a result, grade 3 and 4 students are grouped together. So are grades

students had a perfect score on the CAT/2 test (e.g. 38% for grade 3 students). In wave 2, to reduce the ceiling effects observed in the first wave of the survey, separate versions of the test were created for each school grade. Also note that the response rate in wave 1 for the mathematical component was relatively low (51%).

5 and 6, grades 7 and 8, and grades 9 and 10.[25]

Table 1 shows the number of subsamples we have at our disposal to estimate the impact of the reform by school cycle. We use a total of 11 subsamples: 3 for cycle 1 students (grade 2 only), and 2 for each of the other cycles. We excluded a few subsamples from the original NLSCY data set based on the low response rates and/or the comparability of the scores over time. Further details on our choice is provided in Appendix. The compared subsamples are clearly identified in the tables in which we present our results.

In order to track children over time, we divided longitudinal children into 8 cohorts (see Table 8 in Appendix). Children entering cycle 1 during the same academic year are in the same cohort. Cohort 1 children entered grade 1 or 2 in academic year 1992, and cohort 8 children entered these same grades in 2008. Every two years between 1992 and 2008, with the exception of 2002, a new cohort entered school and was surveyed through the NLSCY. The cohort number given the grade and academic year is specified in Table 1 (columns 3 and 4) in corresponding order.

Table 1 shows that grade 2 students in Québec in cohorts 5, 7 and 8 are considered treated, and may be compared with grade 2 students in cohorts 3 and 4. Note that cohort 4 was surveyed in academic year 1998, when the response rate was relatively low. As a result, we focus on results obtained using cohort 3 (academic year 1996).[26] In grades 3 and 4, treated students are in cohorts 5 and 6, and can be compared with students in cohorts 2 to 4. Given the lower response rate for cohorts 2 and 3, we only present the estimated impact relative to the more recent cohort 4. For grades 5 and 6, treated students are in cohorts 4 and 5. Following the same logic concerning non response, we only compare them with pre-reform students in cohort 3. Cohort 5 is the only cohort including treated students in both grades 7 and 8, since only grade 7 students are treated in cohort 4. Treated students in cohort 5 are compared with pre-reform students in cohorts 2 and 3. As can be seen from

[25] Grade 2 students are not grouped with grade 1 students because, as previously mentioned, grade 1 students are not assessed using the mathematics CAT\2 test.

[26] Results using academic year 1998 are commented in the empirical section, but presented in Appendix (Table 10).

this table, cohort 5 children are treated and observed at different points in time. This allows us to assess the impact of the reform over time on the same children.

From here on, we restrict our attention to observations used to compute the impact of the reform, i.e. all grade 2 students, students in grades 3 to 8 in academic year 2000 to 2006 (except for grades 7 and 8 students in academic year 2004), and students in grades 9 and 10 in academic year 2002, 2004, and 2008.

4.3 Student and family characteristics

Table 2 shows the mean and standard deviation for a number of student and family characteristics. Students attending school before the reform (first two columns) are compared with students attending school after the reform (last two columns). The NLSCY provides a total of 10,268 observations prior to the reform and 19,537 observations after the reform.

The top panel of Table 2 shows that an equal proportion of male and female students are observed prior to and after the reform. The proportions of students per school cycle (grade 2, grades 3 and 4, grades 5 and 6, and so on) depend on the number of cohorts observed for each grade prior to and after the reform. The change in proportions confirms the incidence of the reform on our observed sample displayed in Table 1. For grades 3 to 6, a larger share of students is observed post-reform, while for grades 7 to 10, a larger share of students is observed pre-reform. For about 79% of the students in our sample, a measure of early childhood ability at age 4 and 5 is available, the Peabody Picture Vocabulary Test-Revised (PPVT). This measure is widely used in the literature related to early childhood cognitive development to assess receptive and hearing vocabulary. Table 2 shows that early childhood ability for students observed prior to and after the reform was similar. This implies that changes in students' ability over time may not be attributed to early childhood differences as measured by the PPVT. From Table 2, one can also observe that the trend over school cycles on the CAT/2 test is increasing. As mentioned above, this test is designed to numerically represent the progression of students in mathematics throughout the years. Comparisons of

the mean scores before and after the reform suggest that the overall trend is negative. This may be attributed to an overall trend across Canada and/or a Québec specific effect following the reform. Our empirical strategy outlined above allows us to differentiate between overall trend effects and reform specific effects. Estimated reform effects are presented in the next section.

The bottom panel of Table 2 focuses on family characteristics. Students observed prior to and after the reform share a number of family characteristics. Their family structure, the probability that their mother works and their area of residence are comparable. Household income increases post-reform. In our empirical approach we use household income quartile by year and by region (Québec and RofC), when controlling for confounders. Maternal education is generally higher post-reform.

Although students observed prior to and after the reform 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 Mathematics scores over time and grades

Table 3 shows the mathematical assessment summary statistics for the different possible sets of treated versus non-treated students by school grade. The first panel shows the summary statistics for grade 2 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 first column of the table shows the number of observations for each of the four groups. Children in the RofC observed prior to the reform are labeled Control Before, while those observed after the reform are labeled Control After. Children in Québec observed prior to the reform are labeled Treated Before, while those observed after are labeled Treated After. The second column shows the mean value of the standardized CAT/2 scores by group. The last four columns show the score value at the 25^{th}, 50^{th}, 75^{th} and 90^{th} percentile.

In grade 2 (top panel), the base year (prior to the reform) is always 1996, while the

treatment years are 2000, 2006, and 2008 (from top to bottom). Mean outcomes show that Québec students consistently score higher than students in the RofC on average. Values at the 25^{th}, 50^{th}, 75^{th} and 90^{th} percentile show that this is also true across the distribution. Looking briefly at the evolution of the scores over time, the summary statistics suggest that the scores have been downward trending in both groups (Québec and RofC), but the decrease is more striking in Québec. In 1996, the response rate was below 80%, while it was above 90% in 2000, 2006, and 2008. As mentioned above, in 1996, the tests were still 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 waves 1 to 3 are overestimated, in both Québec and the RofC. If they are overestimated by the same magnitude, our estimates should be unbiased.

In both grades 3-4 and grades 5-6, the base year prior to the reform is 2000. Looking at the mean values, we find a slight increase in the RofC from 2000 to 2002 in both grades 3-4 and 5-6, and a slight decrease in Québec. Now looking at the progression in grades 3-4 from 2000 to 2006, and in grades 5-6 from 2000 to 2004, we observe a slight decrease in the RofC and a much larger decrease in Québec. Similar findings are also observed across the distribution. The fourth panel reveals a similar story for grade 7-8 students (i.e a slight decrease in RofC, and a larger decrease in Québec) whether one looks at 2000 versus 2006 or 2002 versus 2006. 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 versus 2008. Grade 9-10 students in the RofC generally perform better in 2008 (compared with both 2002 and 2004).

Figure 9 shows the differences in mean score between Québec and the RofC over time. The vertical line in each quadrant marks the first school year during which the reform was implemented. In the bottom right figure, there are two vertical lines because the reform was first implemented in grade 7 in academic year 2004, while it was implemented later in

2005 for grade 8. This figure highlights two attractive features of the data: (1) differences between Québec and the RofC were fairly stable prior to the reform, except for grades 7 and 8, and (2) in each grade, mean differences drop following the reform.[27]

To discover whether the instability in the differences in mean outcomes in grade 7-8 were due to the change in characteristics of students and/or the proportion of students in each grade within the grouping, matched samples of students were created. Within each school grade, Québec students in each academic year were matched to Québec students in academic year 2000. The same procedure was applied to students in the RofC. The following matching covariates were included: maternal education dummies, gender, area of residence, household income quartile, and a maternal work dummy. Figure 9 shows the average score differences between Québec and the RofC over time for these matched samples. The trend over time becomes much more stable for grade 7-8 students suggesting that students' characteristics were driving the instability. This may be in part attributed to the rate of non response in academic year 1998 (in wave 3).

In sum, it appears that the reform had negative impacts on the development of mathematical abilities for students in Québec. The following section further validates these results and computes the significance of those differences using standard DID, MDID and CIC.

5 Estimation results

Table 4 presents the empirical results using DID and MDID. Estimated impacts using DID, DID with covariates, and MDID using three matching techniques, suggest that the reform had significant negative effects on mathematical abilities. We first focus on the results using MDID, and then compare these results with the more restrictive DID estimates.

Across all grades, the estimated impacts of the reform are negative and statistically significant. Pre-reform, students in Québec had higher scores in mathematics than students

[27] The trend for grade 9-10 students is not presented as it contains only three comparable observation points (two prior to the reform and one after).

in the RofC. Post-reform, this difference has almost completely vanished. In grade 2, the estimated impacts, all negative, range from 6.2 to 25.4 (14% to 55% of a standard deviation). In grades 3-4, they range from 3.4 to 11.9 (6% to 22% of a std. dev.), while in grades 5-6 they range from 9.2 to 21.4 (16% to 37% of a std. dev.). Finally, in grades 7-8, the estimated effects range from 22.3 to 32.5 (31% to 46% of a std. dev.) and in grades 9-10 they range from 22.8 to 34.6 (25% to 38% of a std. dev). In general, the magnitude of the estimated effects are larger the higher the school grade (both in absolute value and in unit of a standard deviation), but are of comparable magnitude from grade 7 to 10. 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 two years (in grade 4 in 2001, and in grade 5 in 2002) and students in grade 6 had been in the reform for only one year.[28] It is therefore not surprising to find that the estimated effect is smaller in magnitude for that cohort. More surprising are the large effects estimated for grade 2 students (of about 25% to 50% of a standard deviation). Logistic regressions on non-response to the math test reveals that grade 2 students in both waves 2 and 3 had a significantly lower PPVT score even when controlling for the characteristics presented in Table 2. These regressions were estimated on the subset of observations for which the PPVT was available (73% in wave 2, and 90% in wave 3).

Comparing the MDID estimates with the DID estimates, we find that DID estimates (with and without covariates) generally have confidence intervals that considerably overlap with those of the MDID estimates using all three techniques (at 5%). Two main differences are noteworthy. First, in grade 2, for academic year 1996 compared to 2008, the estimated impact is significant using MDID, while it is not using DID with and without covariates. Second, in grades 3-4, the MDID estimators and the DID estimates with covariates are

[28] Table 8 in Appendix provides further details on years spent in the reform given the academic year and the school grade.

smaller in magnitude compared to the standard DID estimates. This is in part due to the change in proportion of grade 3 versus grade 4 students over the waves. For example, in academic year 2000 a fairly comparable number of children in grades 3 and 4 were assessed through the NLSCY in both Québec and the RofC. While this was also the case in academic year 2006 in the RofC, in Québec the proportion of grade 3 students (as opposed to grade 4 students) was largely above 50%. Since the scores of grade 3 and 4 students are not on the same scale, changes in proportions drive changes in mean outcomes, and controlling for these proportions becomes important. DID with covariates and MDID both account for changes in proportions.

In sum, mean effects suggest that the reform had negative effects and these effects are larger the longer the student was treated by the reform. Since adding covariates slightly changes the estimated impacts, we investigate the distributional effects of the reform using the CIC model with covariates. Effectively we first estimate the impact of the covariates on the math scores using ordinary least squares. Then, using the residuals, we estimate the effects using the CIC approach. Since DID with covariates and MDID lead to statistically equivalent results, we assume that controlling for X linearly is not crucial for the results.

Table 5 presents the empirical results using CIC.[29] Table 5 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). The first column shows the mean effect. Columns 3, 5, 7 and 9 present the effects at the 25^{th}, 50^{th}, 75^{th} and 90^{th} percentile and thereby provide an overview of the distributional effect of the reform.

The mean effects assuming conditional independence using CIC with covariates are comparable to that using DID with covariates.[30] Focusing on the estimated impact on cohort 5, we find that the magnitude of the effect is increasing with exposure to the reform (except from grade 2 to grades 3-4). In grade 2, the mean effect is 17.0 (37.0% of a std. dev.), while

[29] 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.

[30] Also note that CIC estimates without covariates are comparable to those with covariates.

it increases from to 15.2 (28.0% of a std. dev.) in grades 3-4, to 19.5 (33.7% of a std. dev.) in grades 5-6, to 23.7 to 29.8 (33.3% to 41.9% of a std. dev.) in grades 7-8, to 26.9 to 43.5 (29.9% to 48.4% of a std. dev.) in grades 9-10. Grade 2 shows a particularly large effect. As mentioned above, these estimates are obtained using the less complete and possibly biased data set of students surveyed in wave 2.[31] Comparison of the estimated impact of treatment on cohort 4 and 5 students in grade 5-6 also support the idea that longer exposure results in higher impact. Cohort 4 students were barely exposed to the reform (1 to 2 years) and the estimated impact of the reform is small and negative, but not statistically different from zero. In contrast, cohort 5 students in grades 5-6 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 cohort 5 (exposed 1 year) with those of cohort 7 (exposed 2 years).

Age at first exposure may also be important, but we only have limited information to assess this possibility. Comparing students spending 1 to 2 years in the reform in grades 5-6 (cohort 4) with those in grade 2 (cohorts 5 and 7), it appears that the reform had a greater impact on younger children since the estimated impacts are negative and significant for them, while it is not different from zero for older students in grades 5-6. This finding needs to be interpreted with caution, as estimated effects on grade 2 students rely on observations with higher non response.

Long term effects may also be different 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 estimator for academic years 1998 and 2008 is small and not different from zero).[32] 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

[31] Table 10 in Appendix present the CIC estimates for grade 2 students using wave 3. Results are generally smaller but remain comparable to those using wave 2.

[32] This can also be observed from Figure 9.

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 at the student level, 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. Colloquially, the province changed a dollar for four quarters at the very high cost of possibly 8 years of lower performance in mathematics and important investments in teacher training. 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. As such, further research is needed to fully understand the long term effects of the reform on a larger diversity of skills.

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. Looking only at the magnitude of the coefficient, it appears that in general students at the 75th percentile have been impacted more negatively, although this difference is generally not significant. 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. Further looking at the top of the distribution (90th percentile), we find negative effects in each cycle, but the estimates are generally not significant. It is thus 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. CIC uses the discreteness of the data to produce upper and lower bounds. Generally, the estimated bounds are fairly tight, which further supports the estimated effects.

In sum, the reform had negative impacts on the mathematical achievement of students in Québec across the entire distribution. These findings are in line with the preliminary findings of Crawford and Snider (2000), evaluating the impact of a more academic approach against

a more contextual approach on a limited sample of 46 students divided in two groups (one treated, the other not). The reform's main objective was to raise the proportion of students who successfully complete their high school education, which means that, indirectly, the main goal of the reform was to raise the achievement of lower performing students. Since mathematical abilities are strongly related to school attainment and labour market outcomes, the evidence presented above suggests that not only did the reform not help these students, but it may actually have been detrimental to them.

5.1 Further evidence from international assessment tests

The NLSCY is not the only source of information providing evidence on the reform. International assessments in which Canada participated can also be used for the analysis.

All provinces in Canada participated in the PISA since 2000. Statistics Canada conducted the survey for the OECD and representative samples of students in each province were selected. The PISA assessments in reading, mathematical and science literacy were administered to 15 year olds across Canada. As a result, it is possible to compare the outcomes of Québec students with those of all other Canadian students, prior to and after the reform (2000, 2003, 2006, versus 2009). The 2009 students are all post-reform students in Québec, while the 2000 to 2006 students are pre-reform students. The global comparable scores over time are presented in Table 6. Statistical differences are measured against 2000 for the reading scores, against 2003 for the mathematical scores, and against 2006 for the science scores.[33]

Canada has been among the top performing countries in PISA over the years, with the provinces of Québec and Alberta scoring highest among the provinces. Québec's performance in reading has decreased over time, when comparing 2009 with 2000, while it has generally been stable over time for all of Canada. Québec's performance in mathematics, and also that of Canada, has been stable over the period, when comparing 2009 with 2003. In science,

[33]The PISA 2000 scores in mathematics, and the 2000 and 2003 scores in science are not comparable to the other scores and are therefore not presented in Table 6.

when comparing 2009 with 2006, there is a slight decrease for almost all provinces (7 points for Québec), but these differences are not statistically significant.

A major downside to the PISA results for the purpose of this analysis is the very low response rate for the province of Québec in 2009 (71%, well below the international satisfactory threshold of 80% set by PISA). A non-response bias analysis conducted by Statistics Canada showed that students in less favorable socioeconomic environments were less likely to participate in PISA and that these students had a statistically lower performance on the provincial reading test (although the difference was small).[34]

A few provinces participated in the Trends in International Mathematics and Science Study (TIMSS). This survey collects data on mathematical and science literacy and is administered to students in grades 4 and 8. The global scores are presented in Table 7. Grade 4 students are post-reform in years 2003 and 2007, and pre-reform in years 1995 and 1999, while grade 8 students are post-reform in year 2007 only, and pre-reform otherwise. In mathematics and sciences, grade 4 students in Québec had a significantly 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, the overall performance of Québec's 4^{th} 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 are compared with results from all previous years: Québec's performance in both mathematics and sciences is going down, while the performance in Ontario is increasing or at worst stable.

Overall, the evidence from these surveys suggests a worsening of Québec's students performance post-reform or at best a stand still. These results are in line with the more detailed results estimated using the NLSCY. As mentioned above, PISA results may be upward biased due to Québec's high non-response rate in 2009. TIMSS results are only partial as the

[34] The PISA 2009 survey also shows a higher participation rate from private school children. These students have an average score of 599 in mathematics compared to 529 for public school students. This effect can be explained by the efficiency of private schools (Lefebvre, Merrigan, and Verstraete, 2011).

only consistent points of comparison are the results from the province of Ontario.

6 Conclusion

We estimated the impact of the Québec school reform on grade 2 to 10 students using math scores provided by the NLSCY. To our knowledge, no formal evidence based evaluation of the reform has been conducted to date.[35]

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 students at all points on the skills distribution. Results based on a small subset of observations, suggest that long run effects may have been null. As such, the reform seems to have failed to meet its primary objective. 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 labour 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 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 drill 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 practices and a more mechanical approach. Pinker (1997) suggests that the poor performance

[35] A research group from Laval University (ERES) has been mandated by Québec's Department of Education to report on the implementation (of cross-curricular competencies), teaching practices and outcomes of high school students. The report is due in 2013, and will rely on data collected since August 2007 (seven years after the beginning of the reform).

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, behavioural 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.[36] We do not measure the impact of the reform on non-cognitive skills, and may be missing part of the benefits (or losses) generated by the reform for high achieving students (and possibly all other students).

Trends in dropout rates across the country between 1990-91 and 2009-10 suggest that, if anything, we may be missing further negative effects. While the overall rate fell from 16.6% to 8.5% in Canada, in Québec it fell from 17.4% to only 11.7% (Statistics Canada, 2010). While Québec had the third highest dropout rate in Canada in 1990-91, it had the highest rate in all of Canada by 2009-10. Clearly, even if social skills were improved, they did not help achieve the reform's primary objective which was to ensure the success of each and every student (MELS, 1999).

While improving non-academic skills may be well placed, the negative effects on academic skills remain worrying. Future research should focus on the long run effects, on both cognitive and non-cognitive skills, across the skills distribution.

[36] As mentioned above, Table 9 in Appendix provides the complete list of competencies and areas of learning.

7 References

Abdulkadiroglu, Atila, Josh Angrist, Sarah Cohodes, Susan Dynarski, Jon Fullerton, Thomas Kane, and Parag Pathak. 2009. "*Informing the Debate: Comparing Boston's Charter, Pilot and Traditional Schools.*" Boston: The Boston Foundation. http://www.tbf.org.

AIR. 2005. "Report on Elementary School Comprehensive School Reform Models." Comprehensive School Reform Quality Center. Washington D.C. American Institutes for Research.

Aladjem, Daniel K., Kerstin Carlson LeFloch, Yu Zhang, Anja Kurki, Andrea Boyle, James E. Taylor, Suzannah Herrmann, Kazuaki Uekawa, Kerri Thomsen, and Olatokunbo Fashola. 2006. "*Models Matter—The Final Report of the National Longitudinal Evaluation of Comprehensive School Reform.*" Washington, DC: American Institutes for Research.

Angrist, Joshua D., and Alan D. Krueger. 1999. "Empirical Strategies in Labor Economics." *Handbook of Labor Economics*, O. Ashenfelter and D. Card, eds. North Holland: Elsevier, chapter 23: 1277-1366.

Athey, Susan, and Guido W. Imben. 2006. "Identification and Inference in Nonlinear Difference-In-Differences Models." *Econometrica* 74(2): 431-97.

Berends, Mark, Susan J. Bodilly, and Sheila Nataraj Kirby. 2002. "Facing the Challenges of Whole-School Reform: New American Schools after a Decade." Santa Monica, CA: RAND Corporation.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119(1): 249-75.

Besley, Timothy J., and Anne Case. 2000. "Unnatural Experiments? Estimating the Incidence of Endogenous Policies." *Economic Journal*, 110(467): F672-94.

Blundell, Richard and Monica Costa Dias. 2009. "Alternative Approaches to

Evaluation in Empirical Microeconomics." *Journal of Human Resources* 44(3): 565–640.

Bifulco, Robert, William Duncombe, and John Yinger. 2005. "Does Whole-School Reform Boost Student Performance? The Case of New York City." *Journal of Policy Analysis and Management* 24(1): 47–72.

Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2003. "Comprehensive school reform and achievement: A meta-analysis." *Review of Educational Research* 73(1): 125–230.

Bulle, Nathalie. 2011. "Comparing OECD educational models through the prism of PISA." Forthcoming. *Comparative Education*.

Bussière, Patrick, Fernando Cartwright, and Tamara Knighton. 2004. "The performance of Canada's youth in mathematics, reading, science and problem solving: 2003 first findings for Canadians aged 15." Statistics Canada. Catalogue no. 81-590-XIE2004001

Bowlby, Jeffery W. and Kathryn McMullan. 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

Crawford, Donald B., and Vicki E. Snider. 2000. "Effective mathematics instruction the importance of curriculum." *Education and Treatment of Children* 23(2): 122-42.

Currie, Janet, and Duncan Thomas. 2001. "Early Test Scores, Socioeconomic Status and Future Outcomes." *Research in Labor Economics* 20: 103-32.

Deke, John and Joshua Haimson. 2006. "Valuing Student Competencies: Which Ones Predict Postsecondary Educational Attainment and Earnings, and for Whom?" *Mathematica Policy Research*, Princeton, NJ. Submitted to: Corporation for the Advancement of Policy Evaluation.

Donald, Stephen G. and Kevin Lang. 2007. "Inference with difference-in-differences and other panel data." *The Review of Economics and Statistics* 89: 221–33.

Ertl, Hubert. 2006. "Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany." *Oxford Review of Education*,

32(5): 619-34.

Fleischman, Howard L., Paul J. Hopstock, Marisa P. Pelczar, and Brooke E. Shelley. 2010. "*Highlights From PISA 2009: Performance of U.S. 15-Year-Old Students in Reading, Mathematics, and Science Literacy in an International Context*" (NCES 2011-004). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office

Fuchs, Thomas and Ludger Woessmann. 2007. "What Accounts for International Differences in Student Performance? A Re-Examination Using PISA Data" *Empirical Economics* 32(2-3): 433-64.

Gleason, Philip, Melissa Clark, Christina Clark Tuttle, and Emily Dwoyer. 2010. "*The Evaluation of Charter School Impacts: Final Report.*" (NCEE 2010-4029). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gonzales, Patrick, Trevor Williams, Leslie Jocelyn, Stephen Roey, David Kastberg, and Summer Brenwald. 2008. "*Highlights From TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context.*" (NCES 2009–001 Revised). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Grek, Sotiria. 2009. "Governing by numbers: the PISA 'effect' in Europe," *Journal of Education Policy* 24(1): 23-37.

Hanushek, Eric A. and Steven G. Rivkin. 2009. "Harming the Best: How Schools Affect the Black-White Achievement Gap." *Journal of Policy Analysis and Management* 28(3): 366–93.

Hanushek, Eric A. and Ludger Woessmann. 2008. "The role of cognitive skills in economic development." *Journal of Economic Literature* 46(3), 607-68.

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

——. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64(2): 605-54.

Heckman, James J. and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review* 91(2): 145–49.

Heckman, James J., Jora Stixrud and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24(3): 411-82.

Hoxby, Caroline M., Sonali Murarka, and Jenny Kang. 2009. "How New York City's charter schools affect achievement, August 2009 Report." Second report in series. Cambridge, MA: New York City Charter Schools Evaluation Project, September.

Ingram, Beth F. and George R. Neumann. 2006. "The returns to skill." *Labour Economics* 13: 35–59.

Klugh, Elgin L., and Kathryn M. Borman. 2006. "Comprehensive School Reform vs. No Child Left Behind." in *Examining Comprehensive School Reform*, ed. Daniel K. Aladjem and Kathryn M. Borman. Washington, DC: Urban Institute Press, 143-177.

Knighton, Tamara, Pierre Brochu, and Tomasz Gluszynski. 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, Vi-Nhuan, Brian M. Stecher, J. R. Lockwood, Laura S. Hamilton, Abby Robyn, Valerie L. Williams, Gery W. Ryan, Kerri A. Kerr, Jose Felipe Martinez, and Stephen P. Klein. 2006. "Improving Mathematics and Science Education: A Longitudinal Investigation of the Relationship between Reform-Oriented Instruction and Student Achievement." Santa Monica, CA: RAND Corporation. www.rand.org/pubs/monographs/MG480.

Lefebvre, Pierre, Philip Merrigan, and Matthieu Verstraete. 2011. "Public subsidies to private schools do make a difference for achievement in mathematics: Longitudinal evidence from Canada." *Economics of Education Review* 30(1): 79-98.

MELS. 1999. "The Education Reform What It's All About." Gouvernement du Québec, Ministère de l'Éducation, du Loisir et du Sport. 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.

MELS. 2003. "Québec Education Program: Secondary Cycle One." Gouvernement du Québec, Ministère de l'Éducation, du Loisir et du Sport.

MELS. 2005. "Education in Quebec. An Overview." Québec: Ministère de l'Éducation. 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.

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

Murnane, Richard J., John B. Willett, Yves Duhaldeborde, and John H. Tyler. 2000. "How important are the cognitive skills of teenagers in predicting subsequent earnings?" *Journal of Policy Analysis and Management* 19(4): 547–68.

Neal, Derek A. and William R. Johnson. 1996. "The role of pre-market factors in black-white differences." *Journal of Political Economy* 104(5): 869–95.

Orland, Martin, Amanda Hoffman, and E. Sidney Vaughn. 2010. "*Evaluation of the Comprehensive School Reform Program Implementation and Outcomes: Fifth-Year Report.*" Washington, D.C.

Pinker, Steven. 1997. "How the Mind Works." New York: W. W. Norton & Compagny.

Rose, Heather, and Julian R. Betts. 2004. "The Effect of High School Courses on Earnings." *Review of Economics and Statistics* 86(2): 497-513.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*. 70(1): 41–55.

Statistics Canada. 2010. "*Trends in dropout rates and the labour market outcomes of young dropouts.*" Education Matters: Insights on Education, Learning and Training in Canada 7(4). Ottawa. www.statcan.gc.ca/pub/81-004-x/2010004/article/11339-eng.htm

Vernez, Georges, Rita Karam, Louis T. Mariano, and Christine DeMartini. 2006. "Evaluating Comprehensive School Reform Models at Scale: Focus on Implementation." Santa Monica, CA: *RAND Corporation*, http://www.rand.org.

——. 2004. "Assessing the implementation of Comprehensive School Reform models." Santa Monica, CA: *RAND Education*, http://www.rand.org.

8 Tables

Table 1: SUBSAMPLES COMPARED

	Academic years		Cohorts	
School grades	Before	After	Before	After
2	1996	2000, 2006, 2008	3	5, 7, 8
3 and 4	2000	2002, 2006	4	5, 6
5 and 6	2000	2002, 2004	3	4, 5
7 and 8	2000, 2002	2006	2, 3	5
9 and 10	2002, 2004	2008	2, 3	5

Note: Shows, for each school cycle, the academic years pre and post reform for which mathematical assessment scores are available and comparable over time, with their corresponding cohort number on the right (columns 3 and 4).

Table 2: Summary statistics

	Before		After	
	Mean	Std. dev.	Mean	Std. dev.
Student characteristics				
male	0.51	0.50	0.50	0.50
school grade				
2	0.13	0.33	0.18	0.38
3 and 4	0.16	0.37	0.26	0.44
5 and 6	0.15	0.36	0.30	0.46
7 and 8	0.30	0.46	0.14	0.35
9 and 10	0.27	0.44	0.12	0.32
ppvt (age 4-5)	100.34	14.65	99.87	15.12
math CAT/2				
grade 2	310.53	45.95	285.42	40.68
grades 3 and 4	367.12	54.25	359.22	51.18
grades 5 and 6	441.13	57.94	434.78	55.09
grades 7 and 8	502.84	71.05	487.44	68.94
grades 9 and 10	589.41	89.92	596.65	87.49
Family characteristics				
family structure				
one parent	0.19	0.39	0.18	0.39
two parents	0.81	0.39	0.81	0.39
household income ('000s)	71.53	53.47	84.32	64.60
mother works (dummy)	0.83	0.37	0.85	0.36
maternal education				
less than secondary	0.12	0.33	0.10	0.30
secondary	0.26	0.44	0.23	0.42
some post-secondary	0.20	0.40	0.14	0.35
college or university	0.41	0.49	0.52	0.50
area of residence				
rural	0.13	0.33	0.13	0.34
urban, $\leqslant$30,000	0.21	0.40	0.16	0.37
urban, 30,000 to 99,999	0.09	0.28	0.10	0.30
urban, 100,000 to 499,999	0.18	0.38	0.17	0.38
urban, $\geqslant$500,000	0.40	0.49	0.44	0.50
Nbr. of weighted obs.	3,909,501		4,534,496	
Nbr. of obs.	10,268		19,537	

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

Table 3: Math score summary statistics by grade

	N	Mean	(Sd)	25th Perc.	50th Perc.	75th Perc.	90th Perc.
GRADE 2							
Year 1996, 2000							
Control, Before	683	313	(47)	286	310	345	368
Control, After	929	289	(42)	264	285	314	345
Treated, Before	135	340	(48)	302	340	402	402
Treated, After	285	298	(39)	269	292	334	361
Year 1996, 2006							
Control, Before	683	313	(47)	286	310	345	368
Control, After	1065	279	(40)	253	274	300	345
Treated, Before	135	340	(48)	302	340	402	402
Treated, After	242	284	(38)	259	274	306	334
Year 1996, 2008							
Control, Before	683	313	(47)	286	310	345	368
Control, After	906	280	(39)	253	274	306	334
Treated, Before	135	340	(48)	302	340	402	402
Treated, After	163	301	(39)	274	292	324	349
GRADES 3-4							
Year 2000, 2002							
Control, Before	1429	358	(52)	317	355	390	425
Control, After	2022	366	(49)	334	359	396	433
Treated, Before	396	391	(53)	361	390	425	458
Treated, After	572	377	(44)	351	375	404	434
Year 2000, 2006							
Control, Before	1429	358	(52)	317	355	390	425
Control, After	3862	347	(52)	310	348	381	416
Treated, Before	396	391	(53)	361	390	425	458
Treated, After	787	353	(53)	317	350	390	425

Note: Shows the summary statistics for each subpopulation by school cycle: grade 2 (top panel) and grades 3-4 (bottom panel).

Table 3: Math score summary statistics by grade (continue)

	N	Mean	(Sd)	25th Perc.	50th Perc.	75th Perc.	90th Perc.
GRADES 5-6							
Year 2000, 2002							
Control, Before	1275	431	(56.)	392	433	468	503
Control, After	1469	440	(51)	407	436	471	507
Treated, Before	307	469	(53)	434	468	510	544
Treated, After	388	465	(53)	425	467	505	532
Year 2000, 2004							
Control, Before	1275	431	(56)	392	433	468	503
Control, After	1956	418	(55)	382	414	456	495
Treated, Before	307	469	(53)	434	468	510	544
Treated, After	528	436	(53)	403	434	468	510
GRADES 7-8							
Year 2000, 2006							
Control, Before	1147	492	(73)	444	492	536	583
Control, After	1834	485	(70)	443	480	529	570
Treated, Before	269	540	(71)	494	529	583	637
Treated, After	518	495	(67)	444	494	540	583
Year 2002, 2006							
Control, Before	1236	492	(68)	448	488	537	577
Control, After	1834	485	(70)	443	480	529	570
Treated, Before	336	524	(58)	490	524	556	594
Treated, After	518	495	(67)	444	494	540	583
GRADES 9-10							
Year 2002, 2008							
Control, Before	864	583	(85)	517	581	642	705
Control, After	1538	596	(88)	530	589	659	717
Treated, Before	184	637	(90)	564	646	704	744
Treated, After	343	599	(86)	530	589	678	703
Year 2004, 2008							
Control, Before	1166	579	(91)	504	574	641	703
Control, After	1538	596	(88)	530	589	659	717
Treated, Before	221	606	(82)	554	605	662	721
Treated, After	343	599	(86)	530	589	678	703

Note: Shows the summary statistics for each subpopulation by school cycle: grades 5-6 (top panel), grades 7-8 (middle panel) and grades 9-10 (bottom panel).

Table 4: DID and MDID Estimated Effect of Treatment on the Treated

	DID		DID with cov.		MDID llr		MDID kernel		MDID neighbor(5)	
	Mean	(Std.err.)	Mean	(Std.err.)	Mean	(Std.err.)	Mean	(Std.err.)	Mean	(Std.err.)
GRADE 2										
AcYear 1996, 2000	-17.321	(8.383)	-21.182	(8.171)	-15.676	(3.492)	-14.129	(4.440)	-13.746	(4.811)
AcYear 1996, 2006	-22.285	(8.365)	-24.307	(8.093)	-22.389	(3.007)	-25.402	(3.577)	-15.624	(3.791)
AcYear 1996, 2008	-6.062	(8.548)	-11.067	(8.372)	-9.368	(4.301)	-10.152	(5.220)	-6.178	(6.537)
GRADES 3-4										
AcYear 2000, 2002	-21.969	(5.025)	-19.636	(4.431)	-11.918	(2.461)	-7.077	(2.848)	-6.566	(2.971)
AcYear 2000, 2006	-27.604	(5.274)	-16.656	(4.870)	-11.053	(3.224)	-3.379	(3.185)	-3.876	(3.552)
GRADES 5-6										
AcYear 2000, 2002	-13.403	(6.230)	-13.428	(5.968)	-13.770	(3.134)	-9.186	(3.778)	-18.042	(4.064)
AcYear 2000, 2004	-20.300	(6.118)	-20.126	(5.855)	-19.098	(3.022)	-20.805	(3.450)	-21.348	(3.835)
GRADES 7-8										
AcYear 2000, 2006	-36.856	(7.814)	-33.428	(7.459)	-30.998	(4.468)	-31.967	(4.967)	-28.530	(5.836)
AcYear 2002, 2006	-22.471	(7.522)	-23.079	(6.695)	-26.288	(4.451)	-32.465	(5.404)	-22.321	(6.121)
GRADES 9-10										
AcYear 2002, 2008	-51.532	(12.413)	-45.121	(10.749)	-23.930	(6.239)	-34.609	(8.289)	-31.444	(7.386)
AcYear 2004, 2008	-23.977	(11.466)	-28.584	(10.730)	-26.394	(6.186)	-29.172	(7.081)	-22.789	(6.850)

Note: Shows the estimated effects of the reform by school cycle. The first column shows the mean effects using standard DID without covariates, while the third column shows the mean effects using standard DID with covariates. Columns 5, 7 and 9 show the mean effects using MDID with local linear regression, kernel and nearest neighbor with 5 neighbors matching respectively. Standard errors are in parentheses. Covariates included are maternal education dummies, gender, area of residence, household income quartile, and maternal work dummy. Results are robust to the inclusion of province of residence dummies or/and age in month on the December 31st, and to the exclusion of the results from students residing in Ontario (Canada's largest province).

Table 5: CIC Estimated Effect of Treatment on the Treated

	Mean	(Std.err.)	25th Perc.	(Std.err.)	50th Perc.	(Std.err.)	75th Perc.	(Std.err.)	90th Perc.	(Std.err.)
GRADE 2										
Years 1996, 2000 (cohort 5)										
DID-level	-20.251	(7.981)	-11.876	(9.516)	-14.876	(12.983)	-40.189	(11.186)	-25.098	(5.810)
CIC disc ci	-17.039	(7.509)	-14.000	(8.967)	-10.000	(12.602)	-28.000	(12.438)	-13.000	(13.362)
CIC disc lower	-17.806	(7.534)	-14.000	(9.083)	-10.000	(12.653)	-28.000	(12.929)	-13.000	(14.122)
CIC disc upper	-16.476	(7.502)	-14.000	(8.918)	-10.000	(12.604)	-25.000	(12.342)	-13.000	(13.360)
Years 1996, 2006 (cohort 7)										
DID-level	-24.006	(8.029)	-12.998	(7.917)	-25.144	(12.607)	-46.544	(11.290)	-40.144	(6.725)
CIC disc ci	-21.215	(7.998)	-8.000	(6.840)	-14.000	(14.843)	-37.000	(14.737)	-31.000	(13.249)
CIC disc lower	-22.020	(7.997)	-9.000	(6.960)	-14.000	(14.809)	-37.000	(14.705)	-39.000	(13.306)
CIC disc upper	-20.582	(7.992)	-8.000	(6.824)	-14.000	(14.891)	-37.000	(14.848)	-28.000	(13.553)
Years 1996, 2008 (cohort 8)										
DID-level	-10.204	(8.163)	-3.870	(8.120)	-10.928	(12.979)	-35.710	(12.092)	-24.382	(13.856)
CIC disc ci	-5.428	(7.918)	2.000	(7.310)	-1.000	(13.356)	-18.000	(13.480)	0.000	(18.143)
CIC disc lower	-6.233	(7.968)	-1.000	(7.490)	-1.000	(13.521)	-18.000	(13.513)	-5.000	(18.342)
CIC disc upper	-4.831	(7.888)	2.000	(7.270)	-1.000	(13.305)	-18.000	(13.634)	0.000	(18.493)
GRADES 3-4										
Years 2000, 2002 (cohort 5)										
DID-level	-19.421	(4.327)	-15.490	(5.541)	-18.490	(4.265)	-16.348	(4.173)	-23.870	(7.072)
CIC disc ci	-15.171	(4.184)	-16.000	(5.913)	-17.000	(4.770)	-11.000	(4.760)	-15.000	(8.098)
CIC disc lower	-15.592	(4.194)	-16.000	(6.044)	-18.000	(4.900)	-11.000	(4.905)	-15.000	(8.194)
CIC disc upper	-14.711	(4.178)	-16.000	(5.923)	-17.000	(4.822)	-11.000	(4.738)	-15.000	(8.083)
Years 2000, 2006 (cohort 6)										
DID-level	-16.379	(4.761)	-23.584	(5.828)	-17.209	(5.291)	-8.584	(4.767)	-10.250	(7.129)
CIC disc ci	-16.722	(5.008)	-21.000	(6.288)	-16.000	(7.254)	-9.000	(5.837)	-11.000	(7.869)
CIC disc lower	-17.215	(5.013)	-22.000	(6.303)	-16.000	(7.354)	-9.000	(5.984)	-11.000	(7.941)
CIC disc upper	-16.144	(5.008)	-21.000	(6.300)	-16.000	(7.196)	-9.000	(5.780)	-9.000	(7.872)

Note: Shows the estimated effect of the treatment on the treated on grade 2 students (top panel) and grades 3 and 4 students (bottom panel). Standard errors are in parentheses.

Table 5: CIC Estimated Effect of Treatment on the Treated (continue)

	Mean	(Std.err.)	25th Perc.	(Std.err.)	50th Perc.	(Std.err.)	75th Perc.	(Std.err.)	90th Perc.	(Std.err.)
GRADES 5-6										
Years 2000, 2002 (cohort 4)										
DID-level	-13.436	(5.973)	-15.487	(4.935)	-9.370	(8.373)	-20.174	(7.725)	-13.243	(8.507)
CIC disc ci	-9.467	(6.052)	-10.000	(5.615)	0.000	(7.799)	-17.000	(10.626)	-16.000	(11.439)
CIC disc lower	-10.025	(6.069)	-10.000	(5.652)	0.000	(7.888)	-17.000	(10.675)	-16.000	(11.341)
CIC disc upper	-8.840	(6.041)	-10.000	(5.649)	0.000	(7.776)	-17.000	(10.788)	-16.000	(11.550)
Years 2000, 2004 (cohort 5)										
DID-level	-20.278	(5.821)	-16.545	(6.239)	-15.576	(7.800)	-30.216	(7.508)	-22.147	(8.657)
CIC disc ci	-19.460	(6.225)	-13.000	(6.830)	-14.000	(9.030)	-35.000	(10.323)	-25.000	(11.801)
CIC disc lower	-20.092	(6.230)	-14.000	(6.977)	-14.000	(9.085)	-35.000	(10.452)	-25.000	(11.893)
CIC disc upper	-18.851	(6.218)	-13.000	(6.766)	-14.000	(9.020)	-32.000	(10.238)	-24.000	(11.768)
GRADES 7-8										
Years 2000, 2006 (cohort 5)										
DID-level	-33.469	(7.417)	-23.378	(11.431)	-39.203	(9.277)	-30.517	(10.009)	-25.190	(9.905)
CIC disc ci	-29.824	(7.363)	-28.000	(11.243)	-39.000	(10.277)	-22.000	(10.740)	-6.000	(14.652)
CIC disc lower	-30.284	(7.382)	-28.000	(11.252)	-39.000	(10.303)	-22.000	(10.792)	-19.000	(14.731)
CIC disc upper	-29.240	(7.344)	-24.000	(11.261)	-39.000	(10.299)	-21.000	(10.699)	-6.000	(14.690)
Years 2002, 2006 (cohort 5)										
DID-level	-22.731	(6.753)	-26.049	(6.432)	-25.031	(7.250)	-10.185	(9.858)	-7.836	(7.353)
CIC disc ci	-23.738	(7.344)	-25.000	(7.577)	-24.000	(8.754)	-12.000	(11.200)	-1.000	(12.784)
CIC disc lower	-24.351	(7.360)	-26.000	(7.547)	-26.000	(8.852)	-12.000	(11.201)	-5.000	(13.128)
CIC disc upper	-23.129	(7.333)	-25.000	(7.606)	-23.000	(8.750)	-12.000	(11.263)	-1.000	(12.682)
GRADES 9-10										
Years 2002 2008 (cohort 5)										
DID-level	-44.127	(10.722)	-49.950	(21.956)	-40.499	(13.442)	-38.417	(13.218)	-21.650	(12.509)
CIC disc ci	-43.508	(11.105)	-66.000	(23.374)	-46.000	(16.775)	-47.000	(15.852)	-24.000	(17.089)
CIC disc lower	-43.907	(11.109)	-66.000	(23.378)	-48.000	(16.876)	-47.000	(15.843)	-24.000	(17.107)
CIC disc upper	-43.127	(11.102)	-65.000	(23.456)	-46.000	(16.780)	-47.000	(15.863)	-24.000	(17.130)
Years 2004, 2008 (cohort 5)										
DID-level	-28.122	(10.541)	-33.689	(14.649)	-29.689	(11.734)	-25.682	(12.070)	-16.131	(16.217)
CIC disc ci	-26.870	(10.287)	-37.000	(14.137)	-24.000	(15.115)	-27.000	(14.215)	-16.000	(15.674)
CIC disc lower	-27.449	(10.277)	-37.000	(14.165)	-24.000	(15.151)	-27.000	(14.140)	-17.000	(15.556)
CIC disc upper	-26.369	(10.300)	-36.000	(14.182)	-24.000	(15.234)	-27.000	(14.403)	-16.000	(15.848)

Note: Shows the estimated effect of the treatment on the treated on grades 5 and 6 students (top panel) grades 7 and 8 (middle panel), and grades 9 and 10 students (bottom panel). Standard errors are in parentheses.

Table 6: Comparion of PISA performance across provinces

	2000		2003		2006		2009	
	Mean	(Std. err.)	Mean	(Std. err.).	Mean	(Std. err.)	Mean	(Std. err.)
Reading (global scores)								
Newfoundland	517	(2.8)	521	(4.9)	514	(5.4)	506	(6.2)
Prince Edward Island	517	(2.4)	495*	(4.4)	497*	(5.1)	486*	(5.5)
Nova Scotia	521	(2.3)	513	(4.4)	505*	(5.7)	516	(5.6)
New Brunswick	501	(1.8)	503	(4.3)	497	(5.0)	499	(5.5)
Québec	**536**	**(3.0)**	**525**	**(5.7)**	**522**	**(6.7)**	**522***	**(5.8)**
Ontario	533	(3.3)	530	(5.1)	534	(6.4)	531	(5.8
Manitoba	529	(3.5)	520	(5.0)	516	(5.7)	495*	(6.1)
Saskatchewan	529	(2.7)	512*	(5.6)	507*	(6.3)	504*	(5.9)
Alberta	550	(3.3)	543	(5.7)	535	(6.1)	533*	(6.7)
British Columbia	538	(2.9)	535	(4.5)	528	(7.1)	525	(6.5)
All of Canada	534	(1.6)	528	(4.1)	527	(5.1)	524	(5.2)
Mathematics (global scores)								
Newfoundland			517	(2.5)	507	(3.1)	503*	(3.4)
Prince Edward Island			500	(2.0)	501	(2.7)	487*	(3.0)
Nova Scotia			515	(2.2)	506	(2.8)	512	(3.0)
New Brunswick			512	(1.8)	506	(2.5)	504*	(3.0)
Québec			**537**	**(4.7)**	**540**	**(4.4)**	**543**	**(3.9)**
Ontario			530	(3.6)	526	(4.0)	526	(3.8)
Manitoba			528	(3.1)	521	(3.6)	501*	(4.1)
Saskatchewan			516	(3.9)	507	(3.7)	506	(3.8)
Alberta			549	(4.3)	530*	(4.0)	529*	(4.8)
British Columbia			538	(2.4)	523*	(4.7)	523*	(5.0)
All of Canada			532	(1.8)	527	(2.4)	527	(2.6)
Science (global scores)								
Newfoundland					526	(2.5)	518	(3.9)
Prince Edward Island					509	(2.7)	495*	(3.5)
Nova Scotia					520	(2.5)	523	(3.7)
New Brunswick					506	(2.3)	501	(3.5)
Québec					**531**	**(4.2)**	**524**	**(4.1)**
Ontario					537	(4.2)	531	(4.2)
Manitoba					523	(3.2)	506*	(4.8)
Saskatchewan					517	(3.6)	513	(4.5)
Alberta					550	(3.8)	545	(4.9)
British Columbia					539	(4.7)	535	(4.8)
All of Canada					534	(2.0)	529	(3.0)

Note: Source: Tamara Knighton, Pierre Brochu, and Tomasz Gluszynski. 2010. Measuring up: Canadian Results of the OECD PISA Study — The Performance of Canada's Youth in Reading, Mathematics and Science — PISA 2009 First Results for Canadians Aged 15 Statistics Canada – Catalogue no. 81-590, no. 4, Table 1.5 and 2.5. Statistically significant difference compared to PISA 2000 for reading, 2003 for mathematics, and 2006 for science are denoted using asteriks. Linkage error are incorporated into the standard error for 2003 and 2006 and 2009. Mathematics reporting scales are directly comparable for PISA 2003 and PISA 2006. In reading literacy, the combined scale was constructed in PISA 2000 and later reading assessments were reported on this scale in PISA 2003 to PISA 2009. Non comparable results are not presented in the above table.

Table 7: Comparion of TIMSS performance across provinces

	Mathematics Achievement Grade 4				Mathematics Achievement Grade 8			
Year	1995	1999	2003	2007	1995	1999	2003	2007
International	500	500	500	500	500	500	500	500
Québec	550	-	506	519	556	566	543	528
Ontario	489	-	511	512	501	517	521	517
Alberta	523	-	-	505	-	-	-	-
British Columbia	-	-	-	505	-	522	-	509
	Science Achievement Grade 4				Science Achievement Grade 8			
Year	1995	1999	2003	2007	1995	1999	2003	2007
International	500	500	500	500	500	500	500	500
Québec	529	-	500	517	510	540	531	507
Ontario	516	-	540	536	496	518	533	536
Alberta	555	-	-	543	-	-	-	-
British Columbia	-	-	-	537	-	542	-	526

Note: Source: Trends in International Mathematics and Science Study (TIMSS), year 1995, 1999, 2003 and 2007.

9 Figures

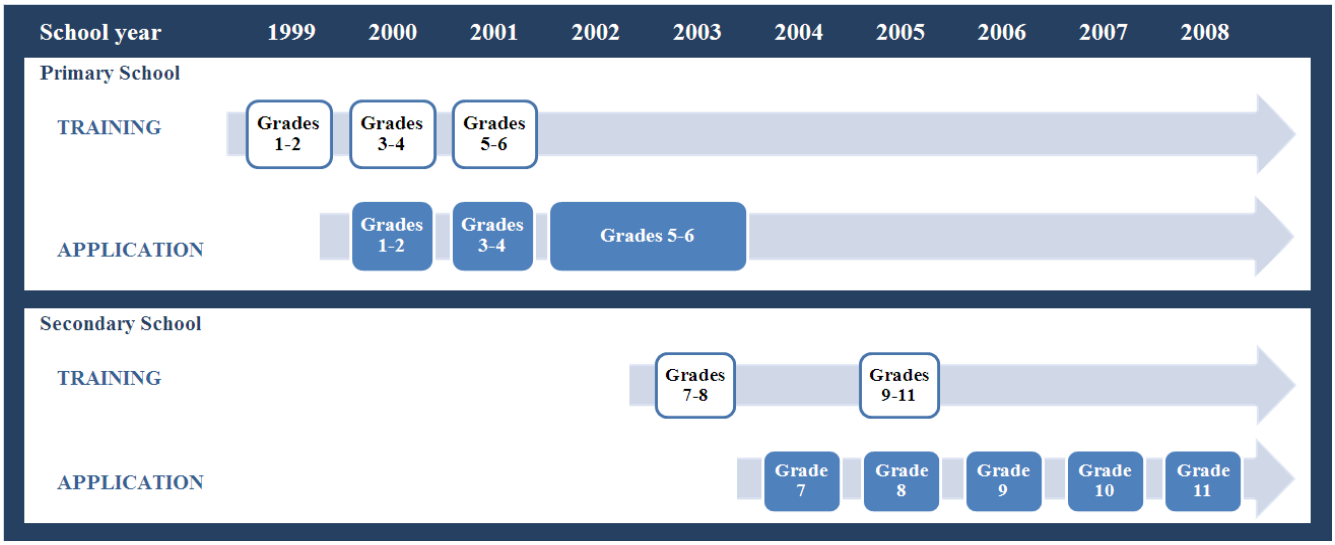

Figure 1: Reform schedule and implementation

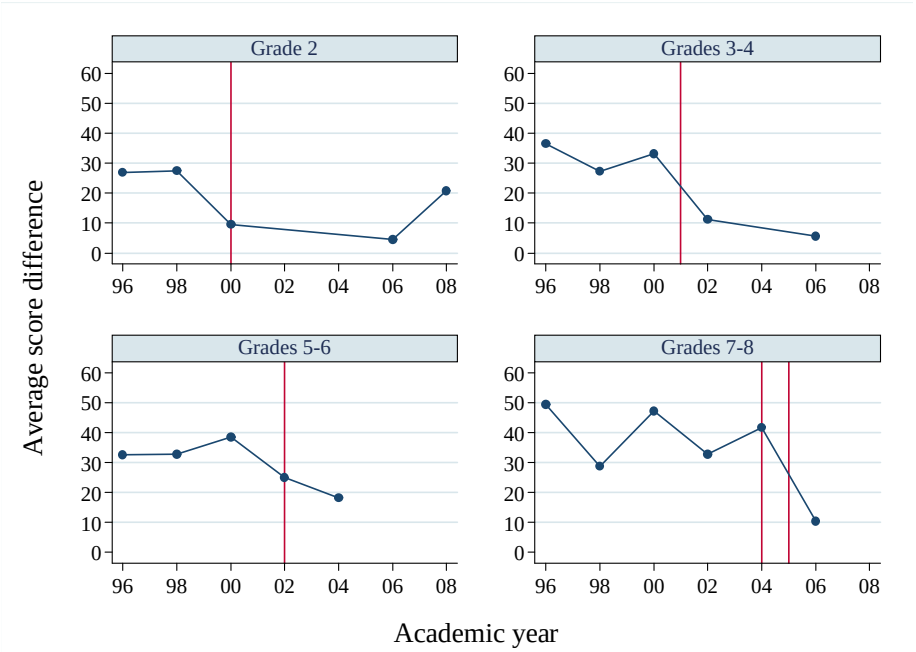

Figure 2: Average score differences: Québec vs RofC

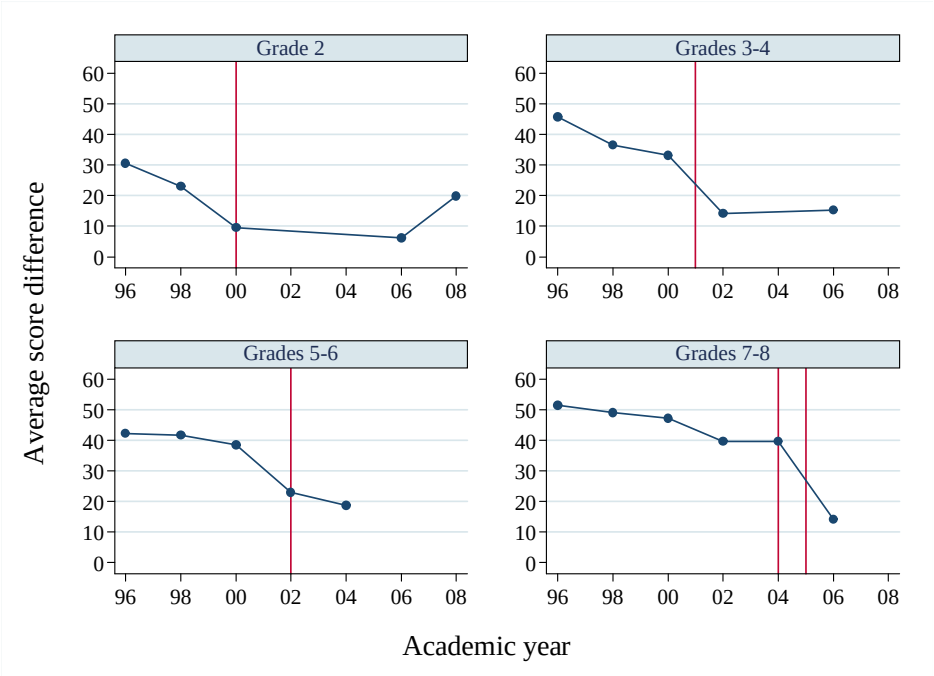

Figure 3: Matched average score differences: Québec vs RofC

10 Appendix

Table 8 provides a more detailed overview of the grades and academic years observed using the NLSCY CAT/2 test. The table covers academic year 1996-97 (wave 2) to academic year 2008-09 (wave 8). Longitudinal children are divided into 8 cohorts, and the grades at which they are observed, given the academic year, is specified in the table.

Table 8: School grades pre and post reform by cohort

	Academic Year							Years in
Cohort	1996	1998	2000	2002	2004	2006	2008	reform
1	5 - 6	7 - 8	9 - 10					0
2	3 - 4	5 - 6	7 - 8	9 - 10				0
3	2	3 - 4	5 - 6	7 - 8	9 - 10			0
4		2	3 - 4	[5 - 6]	[7] 8	[9] 10		0, 1, 2, 4, 6
5			[2]	[3 - 4]	[5 - 6]	[7 - 8]	[9 - 10]	1, 3, 5, 7, 9
6				na	na	[3 - 4]	na	3, 4
7						[2]	na	2
8							[2]	2

Note: Shows the school grades pre and post reform for each of the longitudinal cohort observed (children entering grade 1 or 2 during the same academic year are in the same cohort) as well as the number of years spent in the reform (last column). Boxed grades are under the reform, while unboxed grades are not. Years spent in the reform are presented in corresponding order in the last column of the table.

The incidence of the reform on school levels for which mathematical assessment scores are available is presented in Table 8. Boxed grades observed are under the reform, while unboxed grades observed are prior to the reform.

Years spent in the reform are shown in the last column. Cohorts 1 to 3 students were not impacted by the reform. Cohort 4 was only partially impacted by the reform. Students observed in grades 3, 5, 7 and 9 spent respectively 0, 2, 4 and 6 years in the reform. They were first impacted by the reform in grade 4 of academic year 2001. Students observed in

grades 2, 4, 6, 8 and 10 spent only one year in the reform (i.e. in grade 6). Cohort 5 students observed in grades 1, 3, 5, 7, and 9 were fully impacted by the reform and as a result spent respectively 1, 3, 5, 7 and 9 years in the reform. Cohort 5 students observed in grades 2, 4, 6, 8 and 10, also spent respectively 1, 3, 5, 7 and 9 years in the reform. Grade 1 entrants in this cohort entered school prior to the reform and were only impacted by the reform starting in grade 2 in academic year 2000. Cohorts 6 to 8 were fully impacted by the reform, such that the number of years in the reform equals the grade (e.g. grade 4 students have spent 4 years in the reform).

In sum, the estimated effects of the reform are computed using five different cohorts of treated students (cohorts 4 to 8). While cohorts 4, 6, 7 and 8 provide each only one treated group, cohort 5 provides four treated groups. Pre-reform cohorts surveyed in academic years 2000 and 2002 are preferred because they are more recent and the response rate is higher for both in comparison to those of academic years 1996 and 1998.

Table 9: Competencies and broad areas of learning

Cross-curricular competencies
To use information effectively, and in new contexts
To solve problems using varied and effective strategies
To formulate and exercise appropriate critical judgment
To use creativity in consideration of all elements of the situation
To adopt effective work methods for the task to be performed
To use effectively information and communications technologies
To construct his/her identity
To cooperate with others with appropriate attitudes and behaviours
To communicate appropriately with clarity, coherence, appropriateness and precision
Broad areas of learning
Health and well-being
Career planning and entrepreneurship
Environmental awareness, and consumer rights and responsibilities
Media literacy
Citizenship and community Life

Source: Ministère de l'Éducation, du Loisir et du Sport.

Table 10: Estimated Effect on Grade 2 children (Prior period: Year 1998)

	Mean	(Std.err.)	25th Perc.	(Std.err.)	50th Perc.	(Std.err.)	75th Perc.	(Std.err.)	90th Perc.	(Std.err.)
GRADE 2										
Years 1998, 2000 (cohort 5)										
DID-level	-17.050	(7.210)	-18.895	(9.035)	-16.802	(9.079)	-18.556	(11.929)	-9.786	(8.572)
DID-logs	-16.289	(7.184)	-18.470	(8.804)	-15.607	(8.917)	-16.351	(11.723)	-6.995	(8.468)
CIC disc ci	-16.779	(7.057)	-25.000	(10.076)	-25.000	(8.483)	-18.000	(12.301)	-5.000	(10.072)
CIC disc lower	-17.219	(7.064)	-25.000	(10.103)	-26.000	(8.453)	-19.000	(12.378)	-5.000	(10.382)
CIC disc upper	-16.361	(7.049)	-25.000	(10.341)	-25.000	(8.560)	-18.000	(12.324)	-5.000	(9.999)
Years 1998, 2006 (cohort 7)										
DID-level	-21.425	(7.581)	-19.087	(8.149)	-18.410	(9.053)	-27.460	(11.317)	-25.717	(11.797)
DID-logs	-19.412	(7.345)	-18.795	(7.802)	-16.821	(8.695)	-23.977	(10.799)	-20.829	(11.407)
CIC disc ci	-22.823	(8.410)	-19.000	(9.776)	-20.000	(11.257)	-28.000	(13.955)	-27.000	(16.218)
CIC disc lower	-23.371	(8.410)	-19.000	(9.811)	-21.000	(11.272)	-28.000	(14.066)	-29.000	(16.401)
CIC disc upper	-22.227	(8.415)	-16.000	(9.960)	-19.000	(11.337)	-28.000	(13.971)	-27.000	(16.231)
Years 1998, 2008 (cohort 8)										
DID-level	-9.031	(8.201)	-4.843	(7.482)	-10.043	(9.143)	-16.093	(13.551)	-14.342	(16.725)
DID-logs	-7.212	(7.988)	-4.527	(7.213)	-8.663	(8.881)	-13.131	(13.088)	-10.263	(16.557)
CIC disc ci	-8.930	(8.171)	-5.000	(9.529)	-13.000	(9.277)	-12.000	(13.963)	-3.000	(19.714)
CIC disc lower	-9.471	(8.181)	-6.000	(9.604)	-14.000	(9.307)	-12.000	(13.997)	-3.000	(19.691)
CIC disc upper	-8.365	(8.179)	-5.000	(9.516)	-12.000	(9.268)	-12.000	(14.024)	-3.000	(19.798)

Note: Shows the estimated effect of the treatment on the treated on grade 2 students. The reference period prior to the reform is academic year 1998 (wave 3) in all three cases. Standard errors are in parentheses.