UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Scienze Economiche e Aziendali
“M. Fanno”

CORSO DI LAUREA MAGISTRALE IN

Economics and Finance (curriculum Economics)

TESI DI LAUREA

An assessment of the Gelmini school reform in Italy: a synthetic control approach

RELATORE:

Ch.mo Prof. Lorenzo Rocco

LAUREANDO: Alberto Antonello
MATRICOLA N.: 2002800

ANNO ACCADEMICO 2021-2022

Dichiaro di aver preso visione del "Regolamento antiplagio" approvato dal Consiglio del Dipartimento di Scienze Economiche e Aziendali e, consapevole delle conseguenze derivanti da dichiarazioni mendaci, dichiaro che il presente lavoro non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Dichiaro inoltre che tutte le fonti utilizzate per la realizzazione del presente lavoro, inclusi i materiali digitali, sono state correttamente citate nel corpo del testo e nella sezione 'Riferimenti bibliografici'.

I hereby declare that I have read and understood the "Anti-plagiarism rules and regulations" approved by the Council of the Department of Economics and Management and I am aware of the consequences of making false statements. I declare that this piece of work has not been previously submitted – either fully or partially – for fulfilling the requirements of an academic degree, whether in Italy or abroad. Furthermore, I declare that the references used for this work – including the digital materials – have been appropriately cited and acknowledged in the text and in the section 'References'.

Firma (signature) ....Alberto Antonelli..................

A Simonetta e Attilio

Abstract

I attempt to gauge the impact of the 2008-2010 Gelmini school reform on Italian students' educational achievement. The reform aimed at cutting on educational spending by targeting teaching staff expenditures, as well as boosting the overall efficiency of the education system.

I apply Synthetic Control methods (SCM) to a panel dataset of six PISA international assessments for 25 countries, and carry out a case study of the reform. I find large effects on mathematics performance, but no statistically significant evidence of an impact on reading scores. The inferential strategy based on placebo runs sets the p-value for math treatment effect about the 10% threshold, meaning weak statistical significance. The observed positive effect on math scores may simply be the result of training for Invalsi tests and *teaching to the test*. However, a conservative and economically relevant conclusion can be drawn from my results: the Gelmini school reform did not negatively affect Italian students' achievement in international tests. Contextual questionnaires allow me to provide an interpretation for these results.

In robustness analysis, I experiment with changing the matching period, the predictors, and the donor pool units. Moreover, I directly address the scarcity of pre-intervention observations by merging TIMSS data, and applying a recently developed *penalized* Synthetic Control method.

This is one of the few studies that attempted to quantitatively assess the outcome of the Gelmini reform. Moreover, to the best of my knowledge, my analysis constitutes the first attempt to apply SCM to a merged PISA-TIMSS database, and the first application of the *penalized* SCM to international assessments data.

Keywords: Synthetic control, PISA, TIMSS

Contents

1 Introduction

In 2008, the Berlusconi IV Government passed a series of reforms to the school system aimed at cutting public spending and reducing inefficiencies. The reform package aimed at cutting on educational spending by targeting teaching staff mainly. In order to reduce teaching and non-teaching staff, the Government reduced instruction time in all school cycles, and class sizes were increased at the margins. At the same time, the minister Mariastella Gelmini designed a few interventions to boost the efficiency of the education system: for instance, Invalsi standardized tests were introduced in middle school exit exams. Possibly, the reform represented the biggest change to the Italian education system since the 1960s. Unsurprisingly, the reform triggered heated reactions from the so called "Onda anomala", a protest movement supported by students, teachers and parents concerned with the potential negative effects of the reform on pupils' educational attainment. My thesis attempts to gauge the impact of the 2008-2010 Gelmini school reform on Italian students' educational achievement.

Studying the effects of the reform on educational achievement is extremely relevant from an economic perspective. An early study by Murnane et al. (1995) coupled the assessment of sampled students' achievement in mathematics and reading with a follow up on their adult labor outcomes. They find clear evidence that higher cognitive skills are associated with better employment and wage outcomes. Additionally, Nickell (2004) shows that cross-country variation in wage dispersion is strongly linked to variation in skill dispersion in IALS (International Adult Literacy Survey) data.

Paragraph 5 in Hanushek and Wößmann (2011) provides a detailed literature review on the link between cognitive skills and earnings, underlining the difference between quantitative measures of schooling (e.g. years of school attainment) and qualitative measures (e.g. performance in international assessments). The literature agrees in assigning a large role to institutions and cultural factors in determining skill prices. Measuring educational achievement with cognitive skills rather than school attainment also delivers better results in explaining long-run economic growth[1]. These results appear robust to alternative strategies that address potential endogeneity issues.

Finally, as highlighted by Braga et al. (2011), the more educated are both more likely to be employed and enjoy higher wages, and experience better health and other non-monetary outcomes (social cohesion, political participation, etc.).

I apply Synthetic Control methods (SCM) to a panel dataset of six PISA

[1] Such findings come from studies that applied transformations of the data to link early international assessments; see section 3.

international assessments for 25 countries, and carry out a case study of the reform. This means I create a “synthetic Italy” that tries to approximate true Italy in terms of the outcome of interest and a few covariate predictors. Post-intervention discrepancies between observed and artificial Italy should reflect the effects of the reform, under the identifying assumptions. I find no statistically significant evidence of an impact on reading scores, but large effects on mathematics performance. The inferential strategy sets the p-value for the treatment effect on math scores in a range about the 10% threshold. The staggered implementation of the interventions allows me to attempt an interpretation of these results.

A conservative conclusion that can be drawn from my results is the Gelmini school reform did not negatively affect Italian students’ achievement in international tests. An explanation for such finding may be the partial implementation of the cuts to school resources. The availability of contextual questionnaires allows me to test this hypothesis.

My analysis suffers from the scarcity of pre-intervention data, which impacts the choice of weights for control units and covariate predictors. In robustness analysis, I experiment with changing the matching period, the predictors, and the donor pool units. Moreover, I merged PISA and TIMSS data to increase pre-reform periods, and applied a recently developed variant of SCM that penalizes poor interpolations and improves on the features of standard SCM.

This work is organized as follows: section 2 describes the provisions of the Gelmini reform and summarizes previous literature studying its effects on educational outcomes. Section 3 provides a background on the determinants of educational performance, with a focus on those that were targeted by the reform (i.e. school time, class size, and accountability systems). Section 4 and 5 present the datasets and the methodology, respectively. Section 6 details the results, and section 7 provides robustness analysis. Section 8 concludes.

2 The Gelmini school reform in Italy

2.1 The provisions of the Gelmini reform

Excluding the introduction of middle school and the re-organization of university courses, the Italian school system has not witnessed substantial changes since the 1960s. Volante and Klinger (2021) highlight that, contrary to Estonia, Italy - as well as France and Finland - has witnessed incoherent education policies due to frequent changes in governments. They conclude that poor PISA results have not spurred discussion and reforms. However, a major structural reform did occur.

In 2008, the Berlusconi IV Government (May 2008-November 2011) passed a series of reforms to the school system aimed at cutting public spending and reducing inefficiencies. According to Law 133/08, the "Gelmini reform" - named after minister Mariastella Gelmini - aimed at reducing the net deficit of the public administration to 2.5% of GDP in 2008, 2% of GDP in 2009, 1% of GDP in 2010, 0.1% in 2011. This spending review was formulated as part of a response to the rise of net public debt, with the aim of keeping the debt/GDP ratio under control. The goal set by the Government was to keep it below 103.9% of GDP in 2008, 102.7% in 2009, 100.4% in 2010, and 97.2% in 2011. Contextually, the Government implemented measures to boost GDP, such as investments in innovation, research, and greater use of technology in schools.

Law 133/08 set the following general objectives: (i) to increase the student-teacher ratio by 1 percentage point in the 2009/10-2011/12 period; (ii) to reduce the non-teaching staff by 17% in the 2009-2011 period; (iii) to revise the rules for the composition of classes; (iv) a series of measures (to be defined in subsequent decrees) that would boost the effectiveness and efficiency of the school system. These cuts, together with additional side measures, should have generated reductions in the costs of running the education system equal to 456 million euros in 2009, 1.65 billion in 2010, 2.54 billion in 2011, and 3.19 billion in 2012.

Law 169/08 introduced additional reforms starting in s.y. 2009/10, such as the reform of the single teacher ("maestro unico") in elementary schools, the introduction of a new subject in all public schools - "Cittadinanza e Costituzione" (see Sole24Ore (2008) [1]) -, funding for school buildings maintenance, and provisions for textbooks' editions to reduce spending in instructional material. Subsequent decrees legislated on pre-primary school entry age, grading system, lenght of school hours, and launched a complete revision of high schools since the s.y. 2010/11 (which falls outside the scope of my thesis, as international assessments target pupils younger than 16 years old, and thus they are only marginaly impacted by the reform of high school tracks); they also introduced substantial reforms to the university system.

While these were the general provisions of the Gelmini reform, the relevant legislation is to be found in *implementing decrees* (later presented to the public by Corriere della Sera (2008) and LaRepubblica (2008) [4]). The Decree of the President of the Republic 89/09 reorganized pre-primary, primary and middle schools with the aim to improve the learning opportunities of Italian pupils; in particular, it reduced weekly hours to 30. Prior to that, weekly time included 33 hours of instruction according to Legislative Decree 59/04. The Decree of the President of the Republic 81/09 increased minimum and maximum class sizes to 18 and 27, respectively. The original range for class size was 15-25 according to Ministerial Decree 331/98. These provisions were to be implemented starting in s.y. 2009/10 for grade 6 and subsequently phased in across all grades of middle school.

One of the most relevant interventions for the purposes of this research was the introduction of the Invalsi test within the exam session at the end of the last year of middle school (grade 8). The National Institution for the Evaluation of the Education System (INVALSI) was first set up by Legislative Decree 258/99, and was later reorganized by Legislative Decree 286/04. According to Directive 49/05, Invalsi started the evaluation of pupils' learning achievement (in mathematics and Italian) in elementary schools during s.y. 2005/6 (for grades 2 and 4). The mandatory[2] evaluation of pupils in middle school (in grade 8) only started in s.y. 2007/8, according to Legislative Decree 226/05 and Law 176/07. The ministerial circular 32/08 defines the test's scope, content, administration and evaluation: it states that Invalsi tests aim at complementing the existing evaluation methods and exams, and it sets national standards to be met by schools, thus allowing for greater school autonomy. The circular highlights that the results on the Invalsi test contribute to determining the final exit grade for middle school students, meaning that the Invalsi is a *high-stakes* test. According to Directive 74/08, Invalsi tests would be administered also to pupils in grades 2 and 5 (elementary school) starting in s.y. 2008/9, and to pupils in grade 6 from s.y. 2009/10; prior to that, Invalsi tests in those grades were only administered to samples of schools (Directive 52/07 and Directive 76/09).

The reform generated heated reactions from students, teachers and parents, as documented by many newspaper articles at that time. As documented by LaRepubblica (2008) [1], a first strike by school system workers was organized by independent trade unions (the Italian Cobas) in Rome on October 17th, 2008 ("First No Gelmini Day") - when the Committee on Constitutional Affairs (Italian Senate) passed Decree 137/08. LaRepubblica (2008) [2] records that the protest reached its peak on October 30th, 2008 ("Second No Gelmini Day"), with a strike organized by the major Italian trade union, which involved 80% of teachers across the whole country.

[2] Until s.y. 2006/7, the evaluation of achievements in middle schools was optional.

On the same day, one million people protested in Rome, parading in front of the Ministry of Education (see LaRepubblica (2008) [3]); this coincided with the passing of Law 169/08 (the main provisions of the law were presented to the public by the newspaper article Sole24Ore (2008) [2]). Two other strikes occured on November 14th and December 12th of the year. Protests were so heated that the movement got the name of "Onda Anomala": it was one of the greatest Italian student movements since the 1970s. In 2010, protests reignited with the interventions in the field of university and research (see Sole24Ore (2010)).

Even the then President of the Italian Republic, Giorgio Napolitano, negatively commented the cuts inflicted to the public administration and the school system, stating they were "indiscriminate". Mariastella Gelmini replied that the Government had simply sought to reduce inefficiencies and support the talents (see IlGiornale (2009)).

2.2 Evidence from previous literature

Precisely estimating the effects of the reform in terms of reduced school staff and resources falls outside of the scope of this thesis. For illustrative purposes, I report the main findings in a couple of academic works (see below) and the sketch of the outcomes provided by a newspaper article from LaRepubblica (2011), which gathered the following information:

- 87,400 less teachers in the three-year period 2008/9-2010/1, and additional 19,700 cuts in a.y. 2011/2
- 44,000 less non-teaching staff in the 2010-2012 period
- 25,000 less supply teachers in the 2007/8-2009/10 period
- 10,600 less classes in the 2007/8-2010/1 period, in larger classes and for fewer hours
- overall, a 132 million cut in school system workers and 8 billion euros in resources

Studies evaluating the impact of the reform on students' achievement are scarce, and this work contributes to the understanding of the consequences of Gelmini's intervention. Jahanshahi and Naghavi (2017) exploit Invalsi data (national standardized tests administered to all students in 2nd and 5th grade of elementary school) and the staggered implementation of the reform across grades (the reform was first implemented in grade 1 and later phased in in subsequent grades) to estimate a *triple difference-in-differences* (DDD) model: they find that the reform statistically significantly increased the native-immigrant achievement gap and the gender gap in favor of boys (for both math and Italian tests).

In the book "Bambini che imparano meno? Gli effetti della riforma Gelmini nella scuola primaria", Battistin et al. (2015) tried to provide an overall evaluation of the outcomes of the Gelmini reform for elementary schools (it should be noticed that the focus of this work is specifically on middle schools, instead). They first summarize the main reforms to elementary schools, namely the change in the class formation rule, the abolition of the joint presence of multiple teachers in a class, and the change in minimum and maximum class sizes. Interestingly, they find that the student-teacher ratio and class size did increase and teaching staff was cut in the direction set by the new legislation, but the variation was very modest (class size increased by .27, student-teacher ratio by .67 in the first two years since the reform, and teachers staff was reduced by 9.1%[3]); most likely, the reform did not fully achieve its goals because of exceptions being made for specific schools, and oppositions from parents and trade unions. Moreover, the reform polarized parents' choices towards the extremes of weekly time distribution[4] (classes with less than 30 weekly hours, and classes with more than 40 weekly hours), and the new 24-hours profile (the one that gave the reform the name of "riforma del maestro unico") was chosen only in 0.5% of class formations, implying a negligible effect on pupils[5].

The authors examined administrative sources from the Ministry of Education and the Italian National Institute of Statistics. As a measure of achievement, they use performance in Invalsi mathematics tests during s.y. 2008/9 to 2010/11[6]. They exploit the fact that, while the reform impacted 2nd grade students since s.y. 2009/10, it did not impact 5th year students until s.y. 2012/13; therefore, they can estimate the effect on 2nd grade students using a diff-in-diff strategy. Results point to a 4 percentage points reduction in math scores as a combined result of the Gelmini reform's provisions (reduced teaching staff, increased class sizes, and the abolition of joint teachers' presence[7]).

In a further analaysis, they exploit data from the "Rilevazione sulle Forze di Lavoro" (a labor force survey) to estimate the impact of the reform on pupils' mothers' job outcomes. They find that the reform did not significantly increase their job market participation (except for older mothers), even though the changes pointed to an increase in school time flexibility, with greater consideration given to parents' needs.

[3] The reform aimed at a 17% cut.

[4] The Moratti reform gave more power to parents to influence the organization of school time, meaning that family needs would be taken into consideration by school presidents.

[5] Averaged at the national level, school time did not change significantly.

[6] Invalsi administers questionnaires about contextual variables, too.

[7] The joint teachers' presence explains roughly half of the effect, but its impact is much stronger in the South of Italy with respect to the rest of the country. Estimates based on the same data suggest that teaching staff cuts have much more impact than increased class sizes.

2.3 Education reforms in donor pool units

One of the assumptions behind the application of Synthetic Control methods is that donor pool (see section 5) units did not experience the same treatment as the treated unit, nor did they witness large idiosyncratic shocks that may bias our estimates. Since the donor pool is composed of other countries than Italy, an effort must be made to comprehend their education system and reforms.

Garrouste (2010) discusses the collection of macro data on XX century educational reforms for the 2008/9 SHARELIFE wave of the Survey on Health, Ageing and Retirement (SHARE), which merges micro data on 50+ years old individuals with institutional features of their country of origin. Her database is mostly built on data collected from the Eurydice database, with additions from the Institute for International Education's reports, minor datasets, and other academic papers. She provides an overview of international trends in reforming activity for each level of education (pre-primary, primary, secondary, tertiary), divided by macro-areas, and proceeds to describe the main reforms across most of the countries participanting in SHARELIFE. Finally, she provides an illustrative example of how one could exploit the merged micro and macro data from SHARELIFE to estimate the impact of reforms on compulsory school length.

Braga et al. (2011) created an original dataset of education reforms in 24 countries for the period 1930-2000, using data from Eurydice database, UNESCO country reports, and OECD Education at a Glance; they build tables with level measures of 19 institutional features (duration of compulsory school, age of first tracking, etc.) and temporal variations thereof. They then use this database - together with additional administrative data on school attainment and contextual variables - to estimate the outcome of policy interventions on a quantitative measure of education, with diff-in-diff estimations. Specifically, they exploit the staggered adoption of similar institutional changes across countries, where countries which have not yet implemented the reform constitute a valid counterfactual/control.

In line with these two studies, I carefully reviewed the Eurydice database to explore potential reforms that may cause idiosyncratic shocks in performance for countries included in my donor pool. Additional online sources were consulted for those countries that are not covered by Eurybase. The initial donor pool includes countries that participated in PISA and/or TIMSS waves since 1999, and are in a broad sense comparable to Italy: therefore, one can find European countries (Eastern European, too[8]), USA, Canada, Australia, and New Zealand[9] in it. Then,

[8] For instance, Lavy (2015) finds very similar effects of school time on performance between OECD countries and countries from the former Soviet Bloc.

[9] I originally considered Japand and Korea, but they both experienced large idiosyncratic policy reforms, so that no Asian country is included in my donor pool

in line with suggestions by Abadie (2021), I removed countries that experienced substantial reforms over the 2000-2015 period. I did not exclude countries where the education policy was fluid during the period under consideration. The original donor pool included Australia, Austria, Belgium, Bulgaria, Canada, Croatia, Czech Republic, Denmark, England, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Luxembourg, the Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Spain, Sweden, Switzerland, Turkey, the USA. The final donor pool includes Australia, Austria, Belgium, Bulgaria, Canada, Czech Republic, Denmark, England, Finland, France, Germany, Iceland, Ireland, Israel, Italy, Latvia, Luxembourgh, the Netherlands, New Zealand, Norway, Russia, Spain, Sweden, Switzerland, the USA. Appendix A discusses why specific countries were removed from the original donor pool.

3 Literature Review

3.1 Determinants of educational performance

The main dependent variable used to proxy for educational performance is achievement in international standardized tests (PISA and TIMSS). A cursory review of the literature on the determinants of pupils' performance in those tests is thus required. Moreover, in the context of Synthetic Control methods, an understanding of these determinants is preliminary to later discuss both the choice of predictors and mechanisms. A discussion of international testing features is left to section 4.

International assessments not only provide a powerful measure of human capital (certainly more powerful than a quantitative measure such as school attainment) and cognitive skills, but also make cross-country comparisons possible, allowing for the study of the impact of institutional features on education. They also allow for the study of cross-country heterogeneity in the effects of performance determinants. A major drawback is that they often come as repeated cross-sectional datasets, and student tracking is unfeasible.

Lee and Barro (2001) is one of the first works that try to identify causal relationships between an array of family and school characteristics, and achievement. They exploit a cross-country panel dataset of international assessments (the aggregate level of the analaysis protects against some of the endogeneity issues in regressors), and estimate a standard education production function with country fixed effects, using data from XX century waves of tests administered to 10 and 14 year olds by the International Association for the Evaluation of Educational Achievement and the International Assessment of Educational Progress (additional resource measures are retrieved from UNESCO education indicators). In line with early studies, they find that both parents' income and education strongly affect performance. With the exception of pupil-teacher ratios, the other measures of school resources - teachers' salaries, length of school year, and educational attainment - have a small and not significant effect on test scores (at least when aggregating math, science and reading). The pattern of results is similar when substituting dropout and repetition rates as dependent variables.

While Lee and Barro (2001) use macro-level data aggregated at the national level, Wößmann (2003) uses the student-level TIMSS dataset to estimate an education production function with family background, resources (which, being measured at the aggregate level, are less prone to spillover effects), and institutions as main regressors of interest. Results confirm that family background measures are statistically significant in explaining pupils' performance. In line with early estimates, but difficult to interpret, he finds that higher education expenditure and smaller class sizes seem to reduce performance, possibly because of unaccounted for endogeneity.

Hanushek and Wößmann (2011) closely inspect the classical education production function, which includes students ability (often omitted because it is complex to measure), family background, school resources, and institutions. Most often, data allow to estimate a "snapshot" version of such equation, as information on cumulated resources is often unavailable. Exploiting the PISA database (see section 4), they proceed to estimate the equation, which is able to explain roughly 40% of student-level variation in test performance, suggesting that ability plays a relevant role, and nearly 90% of between-country variation. Overall, they find strong evidence that family background and institutions determine achievement, while evidence on school resources is mixed and not robust. Centrally administered exams (and more generally accountability systems) seem to boost schools and teachers' incentives to deliver higher quality education, while autonomy shows a more complex picture: while school autonomy in setting budget and curricula is negatively related to achievement, autonomy in hiring and firing teachers seems to be beneficial.

Each of the three groups of determinants (family and school resources, institutions) is now explored more in detail. First and foremost, a large literature documents the importante of early child development and family socio-economic background as inputs to educational performance. These are often measured by variables such as parents' income and education, and the number of books at home. Cross-country variation in the degree to which family resources impact achievement can be seen as a proxy for the inequality of educational opportunities and intergenerational mobility. While most studies remain correlational, some of them have played with quasi-experimental designs and instrumental variables (IV) for identification.

Using TIMSS 1995 data, Schütz et al. (2008) regress achievement scores on a proxy for family background resources (the number of books at home, see section 4 below) and controls, and find that the inequality of educational opportunities (i.e. the estimated coefficient on the regressor of interest) varies widely by country, with the US, England and Germany at the top of the distribution, and France, Canada, Belgium and Portugal at the other extreme, implying low levels of educational inequality. Interacting the family background measure with indicators of country-level institutions, they found that (i) early tracking of students increases inequality of opportunities, (ii) that enrollment in pre-primary school has an inverted U-shaped relationship with inequality, and (iii) that the extent of private schools operation fosters an unequal society, but private expenditure in education acts in the opposite direction.

The picture for school inputs is not as clear, and it is often subject of dispute. At the country level, there is no detectable correlation between performance and average expenditure per student. Measures of teaching quality and shortage of instructional material are usually found to be negatively associated with achievement. Class size and instructional time will be discussed more in detail later. Due to potential endogeneity (reverse causality in the resource-achievement relationship), multiple identification strategies have been developed to estimate causal effects, some of which will be explored in the subsection on class size. An illustrative example follows: to estimate the interaction between student and teacher's genders, Schütz et al. (2008) exploit the fact that each student takes different subjects during the same standardized test, allowing to estimate student fixed effects models.

Finally, Hanushek and Wößmann (2011) identify five institutional features that may substantially impact achievement, the first being accountability, reviewed in detail below. The effect of school autonomy is not a-priori clear and depends on whether school's interest points in the direction of increasing pupils' performance: data suggest that students perform better when schools can autonomously decide on processes and personnel, but perform less well when schools decide on their budget and teachers have control over curricula. Results for school private ownership are not robust, likely due to nonrandom selection and general-equilibrium effects: cross-country estimation can help with solving these issues, although omitted variable bias may still be a relevant concern. Interestingly, data suggest the presence of interactions between school accountability, autonomy and private ownership. Early tracking - i.e. channeling students through different school types at an early age - is convincingly found to increase inequality due to selection effects, while no definitive conclusion can be reached regarding the impact on mean performance[10] (see Hanushek and Wößmann (2006)). Given the strong efficiency of interventions early on in students' career (as highlighted above), it is not surprising that the lenght of pre-primary school attendance is positively associated with later performance.

Braga et al. (2011) review additional literature regarding compulsory school length, teacher qualifications, and student financing. By exploiting staggered adoption of education reforms across OECD countries, they conclude that "inclusive policies" and institutions - such as longer compulsory school and increases in tracking age - tend to improve both mean school attainment and equality (raising low- and middle-achieving pupils' scores), while "selective/restrictive reforms" (related to school accountability and teachers' qualification) tend to increase average achievement, but there seems to be no clear impact on dispersion (accountability seems to raise it, while teachers' qualifications points to the opposite direction). At the same time, they find evidence of larger effects among disadvantaged students.

Using empirical calibration methods, Wößmann (2016) obtains time series

[10] Tracking makes teaching easier, as teachers find it more difficult to teach to heterogeneous classes, but pupils risk losing positive peer effects.

of countries' performance in international assessments since 1964. In order to address reverse causality and selection bias, he estimates *first-difference* and *diff-in-diff* models exploiting their panel database. Consistent with previous estimates, he finds that changes in real per capita expenditure do not translate into changes in performance, although evidence from PIAAC database suggests that teachers' quality and practices are beneficial to students' performance. He also exploits country-level variations in school autonomy over time to find that it has a strongly positive impact on performance in OECD countries, but strongly negative in developing ones. In addition to this, he reports findings from a previous study that used historical penetration of the Catholic church as an instrument for the current share of privately operated schools, which seems to positively impact achievement; however, the external validity of such a study is debatable. The relationship between early tracking and inequality is confirmed by Wößmann (2016), once endogeneity is properly addressed.

3.2 School time literature

As discussed in section 2, the reduction in school time (especially in middle schools) was one of the changes brought about by the Gelmini reform. The literature on the role of school time as a determinant of pupils' performance flourished at the beginning of the new millennium, addressing endogeneity effects such as the confounding impact of parents' job market participation. Most of these studies exploit (i) differences in instructional times (e.g. across countries), (ii) reforms, and (iii) specific (targeted) programs. Here, I summarize the findings of a few prominent papers addressing this topic, relying substantially on the list of papers commented by Battistin et al. (2015); it is not intended to be a complete review of the literature on this topic.

Works by Hanushek and Wößmann I presented in previous subsections suggest that resource policies are unlikely to be effective in improving pupils' performance. Lee and Barro (2001) also found non-significant effects of the length of the school term on achievement. The literature focusing on instructional time is not conclusive either, and findings are likely dependent on the specificities of the setting under study. Part of the inconsistencies in results may be related to the fact that interventions during early childhood may deliver higher returns (as suggested by Carneiro and Heckman (2003)).

James-Burdumy et al. (2005) report on the *21th Century Community Learning Centers* program, voted by the US Congress and implemented during the 1990s. This was the first paper providing *experimental* evidence on school time, as low-income elementary and middle schools were randomly chosen to provide academic assistance and recreational activities to their pupils during additional after-school hours. While the program significantly impacted children's supervision, parental outcomes and negative behaviors, it did not affect performance in standardized reading tests and time spent doing homework.

The 2001 Bagrut program was implemented among Israeli low-income high schools in a staggered manner (this features allow for identification of effects, although the program was not implemented randomly). The program granted extra school hours (within smaller classes) to low-achieving students. Lavy and Schlosser (2005) find that the program significantly increased matriculation rates in treated schools, although it was not particularly cost-efficient compared to other interventions.

Pischke (2007) employs an exogenous change in the length of school year for identification. During the 1960s, most German students experienced two "short" years: German states needed to reach uniformity with other European countries regarding the start of school year. German students lost a total of 26 weeks of school out of the reform. The author exploits heterogeneity across cohorts, states and school tracks in exposure to the intervention. Despite the possibility of intervention-specific responses by teachers and schools (to limit the damages), he finds an increase in repetition rates, a reduction in test performance - though short-lived -, and some evidence of changes in track choice.

Marcotte and Hemelt (2008) take a difference perspective to analyse the impact of school time on students: they look at unscheduled school closings due to bad weather in Maryland and find a negative and statistically significant impact on state-level standardized (math and reading) tests, especially when the school closes just before the exam date. To address potential endogeneity of the regressor (e.g. presidents being less prone to school closures because of poor performance of their schools), they exploit data on snow accumulation as an instrumental variable: qualitative results are confirmed.

Bellei (2009) focused on the impact of a 1996 shift from half-day to full school day in Chile; the reform impacted 9th and 10th graders. Schools benefited from large infrastructural investments to implement the change. Schools joined the program in a staggered manner, depending on their ability to provide additional hours to their pupils, allowing for *diff-in-diff* estimation (controlling for covariates that account for treatment-assignment). He concludes that the program had a robust and positive effect on both mean achievement and dispersion of achievement in mathematics and reading.

In the spirit of the No Child Left Behind initiative, Pittsburgh district enacted two after-school tutoring programs during the s.y. 2004-5. Tutoring classes were free and targeted mainly at low-performing pupils. Zimmer et al. (2010) provide an evaluation of the Pittsburgh's program with fixed-effects models and find that it significantly increased math test scores for treated students, but not reading scores.

Similarly to the programs we explored so far, the 2005 Massachussets Expanded Learning Time program provided resources to selected elementary and middle schools to expand school time by at least 300 hours, along with a rescheduling of classes. Boulay et al. (2011) employ a *difference-in-differences* analysis that matches each treated school to one other school with similar characteristics (demographics, achievement, etc.). Findings regarding performance changes are not robust, and teachers' outomes are difficult to interpret. Possibly, this is due to unaccounted for spillovers effects.

Hansen (2011) gathered data on unscheduled school cancellations in Colorando and Maryland that occured before the day in which standardized tests were administered[11]. The regressor of interest is instrumented with a measure of snowfall accumulation, within a 2-sample IV framework (which eliminates the potential selection problems due to data collection in Marcotte and Hemelt (2008)). He complements this analysis with one on test date changes in Minnesota. Weather-related cancellations have a negative and statistically significant impact on math performance. A reduction in school days prior to the test due to test date changes is also negatively related to performance, but the effect is smaller than that for weather cancellations. Reassuringly, the magnitude of the estimates are similar to those in Marcotte and Hemelt (2008), although the estimation tecnique and data differ.

Contrary to most other papers reviewed in this paragraph, which deal with snapshot measures of instructional time, Mandel and Süssmuth (2011) focus specifically on the effects of *cumulated* school time, from grade 1 to test year. They exploit the German national extension of the PISA database (using the first three waves, targeting 15 year olds), and aggregate data at the federal states level. They estimate a classical education production function with many controls and fixed effects, and experiment with different specifications using an *Extreme Bounds Analysis*. They find positive and significant impacts of one additional weekly hour (over nine years of instruction) on math and reading scores, of roughly 15% of an international standard deviation. Most likely, the cumulative nature of their measure drives the magnitude of their estimates compared to previous works.

The consequences of a 2004 school funding reform in Israel is evaluated by Lavy (2012). The reform generated gradual but substantial changes in funding for some elementary schools, although it did not significantly alter overall resources at the country-level. Using data aggregated at the school level to avoid potential selection biases, he estimates an education production function. He then replaces

[11] This strategy is more reliable than studying scheduled variation in length of school year, which may be endogenous (e.g. related to schol budget).

measures of school budget with data on instructional time. Identification comes from school fixed effects, student fixed effects, and IV (with predicted change in funding as instrumental variable): reassuringly, estimates with different strategies are consistent with one another. The author finds that increased instructional budget affects performance in national standardized tests mainly through an increase in school time, although the jump in scores is modest.

Meyer and Van Klaveren (2013) conducted an 11-weeks randomized field experiment where participant students aged 8 to 12 were offered full-day instruction through a voucher system. They estimate an intent-to-treat effect of participation (paired with IV estimation) in order to account for selective non-compliance. They find that the program did not meaningfully affected scores in standardized math and language tests, although the sample size is small.

Rivkin and Schiman (2015) exploit the requirement of PISA 2009 that students must take both math and reading tests to identify the effect of average instruction time: indeed, within school variation across subjects (math, reading) allows the authors to eliminate heterogeneity in general ability and school quality, leaving only subject-specific confounders related to instructional time; to control for selective placement of students into classrooms, they also aggregate data at the school-by-grade-by-subject level. Alternatively, they use within-subject differences across grades to identify the same effect. Their estimates should be taken with caution as some sources of confoundedness remain unaccounted for (even after introducing controls), as pointed out by the authors themselves. Results show a moderate positive effect of instruction time on achievement; there is also evidence of decreasing marginal returns to instruction time and some positive interaction between instruction time and quality of the classroom environment.

Within-student variations across subjects are exploited by Lavy (2015), too. Results show moderate positive and statistically significant effects of teaching time on achievement for PISA 2006 data; on the contrary, OLS estimates appear upwardly biased. Estimates are larger for pupils from disadvantaged families and for developing countries, while estimates for OECD countries and countries from the former Soviet Bloc are very similar to one another. Finally, greater school autonomy in allocating resources (but not in determining the curriculum) and higher degrees of accountability are found to improve the productivity of instructional time. It is remarkable that these estimates are very close to those from Lavy (2012).

Goodman (2014) provides an additional insight on the topic by looking at attendance rates rather than variations in school schedule (instruction time, length of the school year, school closings, etc.). He uses data on demographics, achievement, attendance of students, and school closings in Massachussets. He includes multiple fixed effects and controls to account for potential endogeneity. Absences are significantly associated to lower performance in math and English tests, with the effect being stronger for pupils from disadvantaged contexts. There is no evidence that school closures impact achievement, contrary to what was found by Boulay et al. (2011) and Hansen (2011), suggesting that these studies were mis-identified. He uses two measures of weather conditions to instrument both absences and school closures: estimates remain qualitatively unaffected. The author argues that this pattern of results is due to coordination issues among students, which occur in the presence of absences, but not during school closures (when the entire class is cancelled).

This work contributes to the literature by looking at the impact on achievement of the weekly school hours reduction brought about by the Gelmini reform in middle schools. I employ multiple PISA waves and estimate a *cumulative* effect over the three years of middle school (see section 5). I do not find evidence of lower performance as a result of reduced school time.

3.3 Class size literature

The Gelmini reform changed both the minimum and the maximum number of students per class, causing a variation of class sizes at the extremes of their distribution. Here, I describe the main findings from the literature on the topic. My work contributes to this literature by showing the impact of the Gelmini class size reform on students' performance in international assessments.

Class size - and similarly, student-teacher ratio - is one of the most accurate proxies for school resources: this is easy to see, as teachers' salaries contribute to the majority of a country's educational spending. In the US, substantial federal and states' budget has been devoted to reducing class size since the beginning of the new millennium (for instance, the 1999 federal budget devoted 12 billion US dollars to class size reduction), in an attempt to improve students' achievement; initiatives such as the Project Star have also been developed with the same purposes, but at a reduced scale. Although there is consistent evidence that class sizes did reduce, the economic literature does not agree on its consequences on students' learning.

Interestingly, studies that exploit data at the aggregate level tend to find null to positive effects of reduced class sizes. Lee and Barro (2001) find a negative relationship between pupil-teacher ratio (a proxy for class size) and performance in international assessments. Mandel and Süssmuth (2011), presented in the previous subsection, find that class size has a negative and robust long-run effect on PISA scores across German states, although the magnitude is much smaller than that of instruction time. On the contrary, papers estimating student-level education production functions often find the opposite; for instance, this is the case for

Wößmann (2003). These mixed results are possibly imputable to the lack of an adequate identification strategy. This is problematic, for instance, because class size is determined by stakeholders' choice, which generate estimation biases: school principals can group low achieving students in smaller classes, parents can move to a different district where class sizes are smaller, and policy makers can implement compensatory interventions.

Angrist and Lavy (1999) provide one of the first quasi-experimental evidences on class size: they exploit an Israeli rule on maximum class size (derived from Maimonides' writings, a XII century scholar) as an exogenous cutoff for their regression discontinuity design (RDD). More specifically, they use a *fuzzy* RDD (in practice, this results in an IV strategy), where actual class size is instrumented with predicted class size, based on the formal rule. Their dependent variable is 3rd, 4th and 5th graders' performance in 1991/2 Israeli achievement tests; enrollment and class size data were retrieved from administrative data. They find that larger classes are causally related to lower performance in test performance, across different specifications. There is also evidence that the benefits of smaller classes are more powerful for pupils from disadvantaged contexts. Subsequent studies in the spirit of Angrist and Lavy (1999) use similar rules in other countries (especially Eastern Europe), and do not find large class size effects, although the size of the effect seems to negatively depend on the quality of teaching. Doubts remain as to whether parents exploit the rule to place their children in smaller classes, generating bias.

Hoxby (2000) uses two independent identification strategies to estimate the same effect on Connecticut elementary school students: in the first method, she estimates the random variation in enrollment deriving from random fluctuation of births, and uses it to identify the random variation in class size, which in turn provides an arguably exogenous regressor for pupils' achievement[12]; the second method is very similar to the *regression discontinuity design* employed by Angrist and Lavy (1999), where hitting the class size threshold causes abrupt changes in the number of students per class[13]. While OLS models with controls show a significant negative relationship between class size and achievement, results from the two identificaiton methods do not show evidence of any relevant relationship, even though coefficients are precisely estimated; moreover, the effect does not seem to appear even for low income schools and schools with high percentages of African-American children. The author argues that previous papers that found statisticaly significant

[12] She also experiments with aggregating data at the district level to account for potential spillovers. Finally, she exploits data on kindergarten cohorts instead of enrollment in elementary schools to account for potential relocation choices that occur after parents have observed their children's class size during the first elementary school years. See the original paper for further details on her methodology.

[13] While Angrist and Lavy (1999) use cross-section data, Hoxby (2000) exploits her panel dataset to achieve a more powerful and less biased estimator.

negative effects were possibly biased by the fact that experimental contexts changed the incentives for schools and teachers, while her methodology looks at exogenous variations in class sizes in the natural context where students and teachers act.

In subsequent years, Wößmann and co-authors estimated variants of Hoxby (2000)'s strategy and find consistent results. An additional group of studies, as reviewed by Hanushek and Wößmann (2011), exploits within-student, cross-subjects variation, where students are placed in different classes for different subjects; it should be noticed that unobserved characteristics related to class size that vary by subject may still bias results. No conclusive evidence on class size effects can be drawn from this literature.

3.4 Standardized tests and accountability

As highlighted in section 2, one of the key interventions of the Gelmini school reform consisted in the introduction of the Invalsi test as part of the final examination in grade 8 (last year of middle school in Italy). In this last part of section 3, I briefly review a few studies that explore the topic of accountability as a determinant of educational performance. Accountability is a feature of the educational system that defines the degree to which schools are evaluated by external bodies, mainly through the measurement of students' performance in standardized tests (such as curriculum-based external exit exams).

Although standardized testing was first introduced in the US in mid 19th-century (see Longo (2010)), the interest of the academic community for accountability systems has risen since the enactment of the 2002 *No Child Left Behind* (NCLB) policy in the US, which introduced rewards (e.g. bonuses in terms of extra resources) and penalties (e.g. withdrawal of autonomy, restructuring, closure) for schools based on their performance in state tests, generating pressure on schools to improve their learning outcomes. In the US, the NCLB initiative was the results of a decade-long movement that fostered standard-based examination, but similar policies were implemented in the same period in Latin America, Western Europe, and some developed countries.

The problem with standardized testing is that it places more emphasis on topics that are expected to be on the exam, often leading to the problem of *teaching to the test* and curbing incentives to creative learning. Teaching to the test may also inflate scores without parallel improvements in students' actual understanding of the topic, curbing the ability of the test to measure their true knowledge. Alternative learning practices have been put forward as more effective, such as *inquiry learning*, where the teachers' role is to help students discover knowledge by themselves, boosting their creativity at the same time.

High performance in standardized tests is such a relevant concern for schools that some of them start the school year earlier to give students more time to learn the relevant topics. This is precisely what has been found by Sims (2008) in Wisconsin, where schools adjusted their start date in response to prior performance in state-level exams. He then uses an exogenous reform imposing a later start of school year to find that increasing school time before test date leads to a positive but small increase in achievement.

Lazear (2006) introduced the workhorse model for discussing the introduction of accountability tests. Lazear (2006)'s is a principal-agent model (parents and policy makers acting as principals, students and teachers acting as agents) with some socially-desirable action (preparation for the test) that is subject to monitoring (standardized test) over some space (the topics under examination); when the desirable action is not implemented, the agent is punished if caught. The main insight from such a model is that, when learning is easy (e.g. for high-performing students) and testing learning is cheap, standardized tests may be detrimental to pupils as they tend to narrow teachers' focus on tested topics and reduce students' incentives to learn. However, there are instances in which assessing preparation is difficult and pupils have high learning costs: here, standardized testing is advisable. This is the case, for instance, of schools in disadvantaged contexts. Interestingly, this model can be applied to both students' learning and teachers' choice of curriculum.

In a further effort to emphasize the complexity of the topic, Jennings and Bearak (2014) identify three types of *teaching to the test*[14] and emphasize that all three of them can have positive or negative consequences depending on the context, as suggested by Lazear (2006). Using data on math and language tests in Massachussets, Texas and New York, they find evidence that students adapt their preparation to items that are expected to be on the test based on a review of assessments administered in previous years; in other words, teachers and students react strategically to state tests. Moreover, these results do not seem to be driven by item difficulty (no relationship between difficulty and probability of being tested) or relevance. Finally, there is great cross-state variability in the proportion of state standards that are actually tested on the exam, meaning that the scope for teaching to the test varies by state.

Hanushek and Wößmann (2011) reviews the literature that exploit cross-country variations in accountability systems to find that students under external exit exams regimes and stronger teachers' monitoring of students' performance tend to perform better in international assessments. The problem in identifying causal

[14] (i) Reallocating both between and within subjects to align instruction with state standards; (ii) emphasizing the specific standards predictably represented on state tests; (iii) teaching skills following the same formats in which items appear on state tests.

effects of an accountability system is that it may be correlated with features (especially cultural) of a country which impact performance on their own. Results are robust to the addition of controls for cultural features, and within-country studies (where cultural and language homogeneity is more likely to hold than across countries) deliver consistent findings.

Figlio and Loeb (2011) provide a thorough and specific review of the literature on accountability thus far. The main findings can be summarized as follows:

- the literature finds strong evidence of teaching to the test, i.e. teachers and schools narrowing their curricula and shifting attention away from non-tested topics. This is further supported by the finding that improvements in performance in high-stakes tests (those that have consequences for students) often do not translate to low-stakes tests;
- increasing the scope of testing is costly, but there is high cross-states variability in the proportion of state-level standards that are tested;
- "snapshot" and "growth" approaches to accountability systems provide different incentives (likely dependent on school's performance level) and different degrees of reliability;
- there is modest evidence that both NCLB and similar state- or district-level policies improved students' performance, and the effect is stronger in math than in reading, and more for low performing schools. There is no robust evidence of heterogeneity by race;
- accountability systems generate pressures on teachers and principals, and teachers' turnover is found to increase as a result in low performing schools (with high-quality teachers leaving these schools with more frequency).

Rockoff and Turner (2010) is an exemplary work within this literature in that it studies a new NYC accountability system. The program assigned elementary and middle schools letter grades based on a series of quality and achievement measures[15], generating discontinuities that may impact future school outcomes (not to mention that the system itself allocated rewards and penalties depending on grade achievement). Employing a *regression discontinuity design*, they find that a school receiving a low grade tends to see the subsequent performance of its pupils rise (possibly through *teaching to the test*, but there is also evidence of more time spent on direct instruction and on the use of students' performance data); looking at questionnaires administered as part of the program, it seems that changes occured in low achieving schools have been particularly well received by parents. They do not find evidence of large heterogeneities splitting the sample by schools and students' covariates.

[15] Achievement in test scores, attendance, and evaluations of school environment and quality from questionnaires.

This work contributes to the literature on school accountability because it looks at the impact of the introduction of the Invalsi test in middle schools's final exam (but also in elementary schools). This intervention was explicitly aimed at evaluating each school's added value in terms of improved pupils' achievement, potentially allowing for comparability across schools.

3.5 Synthetic Control Methods in the education literature

Cordero et al. (2018) provide a literature review of 66 studies focusing on the impact of school reforms on pupils' performance, as assessed by large-scale international assessments. The paper highlights that international assessments are handy as they allow the researcher to study reforms that only occur at the aggregate level, and discusses the main sources of endogeneity. The main objective of this work, though, is to provide an exaustive list of empirical methods that can be used to study interventions on education exploiting PISA-like datasets: IV, *regression discontinuity designs*, *difference-in-differences* analyses, and *propensity score matching* are first reviewed in their basic elements; then, a list of applications is provided, with tentative conclusions regarding institutional features such as early tracking, class size, private schools, etc. Not surprisingly, Synthetic Control methods are providing a new avenue of empirical analyses and still do not make into this list.

The application of Synthetic Control methods to the study of educational reforms - especially to international assessments data - has already been explored by some literature (sometimes paired with *difference-in-differences*), but these attempts are all but thorough and conclusive. As highlighted by Johnson (2013), educational achievement data are usually more prone to measurement error than standard economic variables, reducing the power of SCM. Interestingly, though, SCM require the researcher to aggregate data (e.g. at the national level): such aggregation makes international assessments panel data, while they are only repeated cross-sections at the micro-level.

Belot and Vandenberghe (2011) studied the impact of a 2001 reform that reintroduced grade retention in French Belgium: they use PISA scores for all three domains (mathematics, reading and science) over the 2000, 2003 and 2006 (only the 2000 wave is pre-reform) waves to build a synthetic control for Belgium. Two major drawbacks of their analysis should be highighted: first, the availability of only one point in time prior to the intervention, which impairs the credibility of the synthetic control; second, they exploit all other PISA participants as donor pool, regardless of whether all these countries are really comparable (see interpolation bias below), and whether some of them experienced reforms that may have impacted students' performance. Interestingly, they employ standard sample-based statistical inference to explore significance: this is not the best practice in SCM, as it does not take into account the non-random assignment mechanism and clustering at the national level (this was a problem even in earlier case study analyses).

Anghel et al. (2015) applied SCM to PISA data in order to measure the effects of the introduction of a low-stakes standard external examination in the Madrid region since s.y. 2004/5, where other 15 Spanish regions served as donor pool. The 2000 and 2003 waves represent the pre-intervention period, 2006 and 2009 the post-intervention period. Three considerations complicate the validity of their analysis: first, they exploit PISA 2000 mathematics wave, which is notoriously not comparable with subsequent waves; second, the pre-intervention matching is far from perfect; finally, they do not carry out inferential analysis with SCM, which keeps the author from drawing definitive conclusions.

Spanish regions' performance in PISA waves is once again at the basis of a SC analysis in Beneito and Vicente-Chirivella (2020), in particular the waves from 2006 and 2018. The authors are interested in analysing the consequences of new laws banning mobile phones in schools in 2014/2015. It is reassuring that the pre-intervention fit is nearly perfect. It is not clear why the inferencial analysis is not well-addressed in the text (results are described as statistically significant, at least at 10% level), and not shown graphically.

A recent work by Soh et al. (2021) studies the impact of a 2003 change in the language of instruction in Malaysia on teachers and on students' performance, with particular attention given to heterogeneous effects. They employ the 1999, 2003, 2007 and 2011 waves of TIMSS (both for mathematics and science[16]), the former two serving as pre-intervention periods. Although the authors made sure that donor pool countries were not affected by language policy changes, these countries may have been affected by other idiosyncratic shocks that might confound results; moreover, some of these countries (e.g. England, Italy, Morocco, etc.) are not really comparable to Malaysia, increasing the risk of interpolation bias. Not surprisingly, countries such as Slovenia and Romania end up receiving positive weights, although their comparability with Malaysia is debatable[17]. Pre-reform matching is convincingly accurate and the post-reform drop in performance is indeed large, but inference relies on a very small number of countries (between 8 and 9).

This work improves on previous SCM literature because it employs state-of-the-art Synthetic Control methods and inferential analysis, and it excludes countries that are either not comparable to Italy, or they were hit by large idiosyncratic shocks that may confound the results; moreover, effort was put into expanding pre-

[16] Reading is not tested by TIMSS.

[17] Abadie et al. (2010) shows that increasing the size of the donor pool may lead to overfitting, especially when this new units differ greatly from the treated unit in terms of unobserved factors.

intervention data to the fullest, using both PISA and TIMSS datasets. Indeed, I followed the suggestion, posed by Abadie (2021) and Johnson (2013) (the former in the more general context of SCM, the latter in the context of education assessments studies), to include as many pre-intervention periods as possible. Indeed, Abadie et al. (2010) proves that the bias of the SC estimator is bounded by a function that goes to 0 as the number of pre-intervention periods rises. In other words, with few intervention periods, the SC may approximate pre-reform outcomes very well without matching unobservable confounders appropriately, leading to a bias (this is a form of *overfitting*). See section 5 for further discussion on this point.

4 Data

4.1 PISA and TIMSS datasets

Countries around the world started arranging international assessments since at least mid-1960s, with the purpose of testing their pupils on a common set of questions and make international comparisons possible. The 1964 First International Mathematics Study (FIMS) was participated by 12 countries. The International Association for the Evaluation of Educational Achievement (IEA) had gathered academics to develop math, science and reading tests. Since the 1990s, there are three major testing programs: the OECD Programme for International Student Assessment (PISA), the IEA's Trends in International Mathematics and Science Study (TIMSS), and the IEA's Progress in International Reading Literacy Study (PIRLS). I will use the first two databases for my analyses. Table 2.1 in Hanushek and Wößmann (2011) provides a summary of international testing occasions until 2007. They also provide a detailed overview of national and regional tests.

For my empirical analysis, I used aggregated data from the 2000, 2003, 2006, 2009, 2012, 2015 waves of PISA. As noted by the OECD Director for the Directorate of Education and Skills, PISA has become the "world's premier yardstick for comparing quality, equity and efficiency in learning outcomes across countries, and an influential force for education reform" (Schleicher (2019). Additionally, I experimented with merging the 1999, 2003, 2007 waves of TIMSS database. I did not include the most recent wave of PISA (PISA 2018) as many additional countries should be dropped due to recently implemented reforms. Moreover, I did not include more recent waves of TIMSS as the anchoring based on TIMSS 2003 would not be credible for those waves; moreover, the rationale for merging the TIMSS dataset to PISA is to increase pre-reform periods[18], while increasing post-reform ones is not a priority.

The target population for PISA is 15-year-old students, while TIMSS targets 8th grade students (mostly 13 year olds)[19]. At this age in most OECD countries, students are approaching the end of their compulsory schooling. Furthermore, part of the target population is attending lower secondary school, while the other part is attending upper secondary school (especially in PISA). All participating students take a 2-hours tests, followed by a 30-minutes questionnaire. 20-minutes questionnaires are administered to school principals.

Students are tested on three domains in PISA: mathematics, science and reading. TIMSS tests students in mathematics and science (not reading). I focus on

[18] Abadie (2021) explains why adding more pre-intervention periods is much more crucial than adding post-intervention ones.

[19] Since 1999, TIMSS targets 4th graders, too.

mathematics and reading because science tests became comparable only since the 2006 PISA wave. Mathematics became comparable across PISA waves since the 2003 wave[20]. For this reason, I did not use PISA 2000 mathematics data for my estimates. TIMSS data are comparable since the 1995 wave, but I only employ data since 1999 due to the low number of donor pool countries participating in TIMSS 1995. International tests are cross-sectional, meaning that there is no possibility to follow the progression of individual students over time. However, in an effort to achieve cross-country comparability, international tests administer a common set of questions to all participating countries.

While PISA increased participation from 32 countries/regions in 2000 to 72 in 2015, TIMSS went from 38 in 1999 to 48 in 2007 (last wave employed in this study). Participation was voluntary (as for all international assessments), so that the set of participating countries slightly differed over time. In 2015, the number of students per country oscillated between 3,000 and 20,000.

PISA 2015 data were downloaded from `https://www.oecd.org/pisa/data/`, while TIMSS data from `https://timssandpirls.bc.edu/databases-landing.html`. PISA waves before 2015 are not available in a format that is readable by R with `intsvy` package; therefore, I downloaded those data from `https://github.com/pbiecek`.

PISA is designed to provide summary statistics on the population of interest in each country. It does not provide optimal statistics at the student level. Aggregating data at the national level is thus one of the ways a researcher can exploit the database to its fullest.

For the sake of practicality, a two-stage sampling procedure is used in PISA. First, a sample of schools is selected from a list of all the schools in which 15-year olds are enrolled. Then, a simple random sample of students is drawn from within the selected schools. In the second stage, 35 students per school are drawn. TIMSS also uses a two-stage procedure, but once schools have been selected, one entire classroom per grade is randomly picked. PISA requires a minimal student participation rate of 80% in order to limit the size of the bias due to non-response.

Differences in school size is a relevant phenomenon to consider; for example, schools in urban settings tend to enroll more students than schools in rural settings do. Even though, in theory, all schools have the same probability of being drawn, the probability of drawing a certain student differs among schools due to differences in school size. To overcome this problem, schools are actually not drawn with equal probability; on the contrary, they are selected with probabilities proportional to

[20] Scores are standardized to have a mean of 500 and a standard deviation of 100, but such standardization was made in 2003 for math and 2006 for science. The reason why this happened is that PISA has a diffent dominant domain each wave.

their size (larger schools have higher selection probability than smaller ones). This procedure should guarantee that each student has the same selection probability; however, students' data still have to be weighted due to (1) missampling of some strata of the population, (2) lack of accuracy in the measurement of school size and (3) adjustments for student non-response.

Also, because students cannot be considered as independent observations (due to the two-stage sampling[21]), a replication method is suggested for calculating unbiased variances; indeed, with such a complex sampling design, there is no easy formula for computing variances or even means, so that computational intensity needs to be used for retrieving those statistics. In particular, each student is assigned 80 replicate weights (calculated generating 80 replicate samples) according to a Balanced Repeated Replication (BRR) method, in its Fay's variant (with a deflating factor K of 0.5). The statistic of interest will, thus, be computed on the whole sample and then again on each replicate. The comparison between the whole sample statistic and each of the replicate statistics will provide an estimate of its sampling variance. The sampling variance will be computed as:

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-K)^2} \sum_{i=1}^{G} (\hat{\theta}_{(i)} - \hat{\theta})^2$$

where $\hat{\theta}$ is the statistic computed on the whole sample and $\hat{\theta}_{(i)}$ is the statistic computed on the replicate i.

A description of how performance in PISA test is computed is also needed. Performance is not simply represented as the percentage of correct answers, as differences in items' difficulty affect raw performance, making students' raw performance not comparable when assigned tests with different sets of questions[22]: PISA applies the Rasch model, which estimates a student's ability based on both correct answers and items' difficulty. This means that final scores are represented by weighted averages of the correct responses to all questions, with the difficulty of the item used as weight. Items' difficulty is calibrated through a complex process that generates a relative scale of difficulties (a continuum of difficulties) using an items' anchoring process: in other words, the (relative) difficulty of an item results from the comparison with all the other items, where the share of students who manage to get the item right is considered.

Finally, we must take into consideration that PISA database reports student performance through *plausible values* (PVs). This means that posterior distribu-

[21] For instance, notice that students from the same school are not independent, because they likely come from similar backgrounds.

[22] In PISA, students' are assigned different booklets, in order to cover as many topics as possible, and at the same time guarantee tests of adequate length.

tions of students' latent ability are computed around the reported values (i.e. the actual result in the test); then, a series of random values are drawn from the posterior distribution and assigned to the observation. This is done in order to build a continuum from discontinuous variables and compute an unbiased estimate of students' ability. Using this methodology, the researcher also accounts for test unreliability in measuring such ability[23] (imputation variance). In PISA 2015, 10 plausible values are drawn for each student; they were 5 before 2015 and in TIMSS waves.

Population performance statistics are first estimated using each of the PVs. Then, the reported population statistic is the average of these estimates:

$$\theta = \frac{1}{M}\sum_{i=1}^{M}\theta_i$$

The uncertainty in the estimation of the latent variable (i.e. students' ability) is computed as:

$$U_M = \frac{1}{M-1}\sum_{i=1}^{M}(\theta_i - \theta)^2$$

Though other methods (e.g. using only one of the plausible values, or averaging PVs at the student level) give unbiased estimates when computing means, the use of PVs as just described is necessary in that it provides estimates of variances closest to the population value. In particular, the final variance will be computed as:

$$V = \sigma^2_{(\hat{\theta})} + \left(1 + \frac{1}{M}\right)U_M$$

The PISA data analysis manuals contain a much more detailed description of all the relevant features of the PISA dataset, with examples.

Since PISA is the true focus of my analysis due to richer data made available by OECD for the countries in my donor pool, I relied on the user guide provided by Fishbein et al. (2021) for TIMSS 2019[24] as a source of information on TIMSS database. Many more details can be found in TIMSS technical reports. Similarly to PISA, the correct computation of country-level achievement needs to account for 5 plausible values and complex standard errors (which include both sampling variance and imputation variance). In particular, correct standard errors are computed using a jackknife repeated replication method (JRR). Both school and teachers' samples were designed to optimize student samples, so that school and teachers' variables

[23] Education measures are particularly prone to measurement error due to the broad concepts that they measure and specific conditions of testing day.

[24] The User Guide also provides an introduction to the IDB Analyzer, a user-friendly software for the analysis of TIMSS data, using SPSS or SAS.

should be treated as students' attributes.

I used the `instsvy` R package to compute unbiased country means of scores and predictors. Caro and Biecek (2017) describes the main features of the package and how statistics are computed, in line with the complexities of the sampling procedure and scoring system.

A final and more general remark should be stated regarding the external validity of estimates based on international assessments. Volante and Klinger (2021) note that, although developed cooperatively, PISA and TIMSS datasets are based on the untested assumption that they reflect similar constructs across widely different cultural contexts. This work focuses on countries that are, in a broad sense, comparable to Italy. These nations are relatively similar with regards to educational institutions, preventing vast cross-country cultural differences.

4.2 Merging PISA and TIMSS

Hanushek and Wößmann (2011) notice that PISA and TIMSS are strongly correlated at the aggregate (country) level: for intance, although PISA tests 15 year olds and TIMSS 8th graders, the correlation between PISA 2003 and TIMSS 2003 among the 19 countries participating in both tests was 0.87 in mathematics (0.97 in science)[25].

A potential concern when merging the two databases is that, while TIMSS is based on common elements of primary and secondary school curricula, PISA measures pupils' ability to apply knowledge and skills and focuses on lifelong learning. Moreover, PISA and TIMSS normalize scores so that they have a mean of 500 and a standard deviation of 100, but participating countries change across waves and across tests, so that they are not perfectly comparable. In particular, PISA focuses on OECD countries, while TIMSS countries are more heterogeneous. Hanushek and Wößmann (2011) concludes that the similarities in the design and the high correlation between the two suggest they are testing a common dimension of skills (human capital). Therefore, they encourage merging the two (after some form of transformation) in order to develop comparable performance indicators. Cross country differences in sample selectivity, which has been criticized by part of the literature, can only bias results if it is systematically correlated with the error term of the equation under estimation (there is evidence that this is not often the case); otherwise, it simply introduces classical measurement error.

While PISA and TIMSS use a psychometric approach to calibrate comparability across countries and over time (e.g. using the Item Response Theory, which

[25] The correlation between PISA 2012 and TIMSS 2011 among the 28 countries participating in both tests was 0.94 in mathematics (0.93 in science).

weights questions by revealed difficulty), comparisons across different tests must be based on empirical calibrations - that is, based on information on overall distributions of scores. A few studies have tried to merge PISA and TIMSS using empirical calibration, but none of their strategy is really suitable in my case. While Ammermüller (2005) and Hanushek and Wößmann (2006) simply use scores as they are without transformation, Schneeweis (2011) and Wößmann et al. (2015) equate the distribution of scores for a common set of countries, but this is a valid strategy as long as they need to tranform just one or a couple of adjacent waves.

Hanushek and Wößmann (2012) merge 1994-2003 international assessments to study the impact of cognitive skills (human capital) on long-run economic growth. They rescale all scores using the US scores in the National Assessment on Educational Progress (NAEP) as a benchmark and assuming constant variation in scores among a group of already developed OECD countries (OECD Standardization Group, or OSG); indeed, the US participated in all international assessments and NAEP is the only assessment that is comparable over time since its first administration in 1969. The details of their procedure are explained in the appendix to their paper.

As noted by Altinok et al. (2014), a major issue with Hanushek and Wößmann (2012)'s methodology is that they do not prove their assumption of stability in the variation of scores across OSG countries; moreover, their strategy makes the rescaling of scores for countries far from OSG average less reliable; finally, they do not include regional assessments (RSATs). Altinok et al. (2014) propose a new combined methodology to put all international assessments on the same scale, simultaneously allowing for tracking trends over time. For IEA assessments prior to 1995, they use the same anchoring procedure that exploits US performance in NAEP. For post-1995 period, the methodology is unfortunately poorly explained (and contradictory): one reads that PISA and NAEP are used for anchoring PISA, that TIMSS is used to predict PISA scores, and TIMSS (and PISA) trends are maintained, but the practical steps of its implementation are unclear.[26]. Their database may be used in the future to estimate synthetic controls going back to as far as the 1960s; however, it would be much more difficult to obtain comparable data on predictors. Wößmann (2016) puts 1964-2003 tests on a common scale using similar procedures as those seen above, and shows that cross-country variation is much larger than within-country variaton over time, although some countries experience substantial improvements or declines over the considered period.

To merge PISA and TIMSS, I relied on the fact that both assessments tested pupils in 2003. Therefore, I make the assumption that the scores should be the same

[26] For instance, they state that TIMSS data is not anchored anymore to NAEP, but still requires an adjustment; yet, they do not explain how this adjustment would work.

in that year (both measure cognitive skills, but see the discussion on comparability in this subsection): therefore, for each country, I compute a coefficient such that multiplying TIMSS 2003 score by it returns the PISA 2003 score; finally, I multiply TIMSS 1999 and 2003 scores by that coefficient. In other words, I anchored TIMSS 2003 to PISA 2003. By visually inspecting data, this methodology seems to preserve natural (upward, downward, or flat) trends in aggregated performance, but it does not do well far from the anchoring period (PISA 2015 and TIMSS 2015 have a correlation in the order of 0.70 aftern transformation), and this is why I keep PISA 2015 and discard TIMSS 2015 data (as well as TIMSS 2011).

Using this procedure, I get a final database with 25 countries: Australia, Austria, Belgium, Bulgaria, Canada, Czech Republic, Denmark, England, Finland, France, Germany, Iceland, Ireland, Israel, Italy, Latvia, Luxembourg, Netherlands, New Zealand, Norway, Russia, Spain, Sweden, Switzerland, USA. All of them participated in all relevant PISA waves, while only 13 of them participated in TIMSS and could be anchored to PISA. These countries are Australia, Belgium, Bulgaria, England, Israel, Italy, Latvia, Netherland, New Zealand, Norway, Russia, Sweden, USA.

A few papers highlight the differences between international assessments and are more hesitant to suggest comparability between PISA and TIMSS. For instance, Brown et al. (2005) notice that the two international assessments differ in non-response rate and proportion of multiple choice questions, as well as other characteristics I have already mentioned above. Once they have aggregated data by taking z-scores (which preserves both ranking and correlations over time) or simply looking at country rankings, they explore the consistency of results across different tests. They find encouraging correlations between PISA and TIMSS results when looking at measures of central tendency (median), but not as much when looking at measures of dispersion. Moreover, correlations are larger within surveys for different subject than between surveys for the same subject. Considering only children of similar age across surveys does not significantly raise cross-survey correlations. Sampling error is a much more relevant concern for measures of dispersion than for the median (completely eliminating sampling error would only increase their correlations by 0.01 to 0.02, according to their calculations). Finally, they considered choices made by organizers relating to the specificities of the item response model (IR) employed by each survey: all surveys apply unidimensional rather than multidimensional IR models, meaning that they make the underlying assumption that higher ability students are invariably more likely to give a correct answer (not allowing for different groups of students to be better at specific types of questions); however, PISA uses a *one-parameter* model, while TIMSS a *three-parameter* model, meaning that the latter allows for pupils' guessing and the former does not. Their estimates[27] point to high robustness to the choice of model for measures of central tendency, not so for measures of dispersion. Results are more consistent when considering only developed OECD countries. In sum, estimates based on means among similar (mostly OECD) countries should not be largely biased by surveys' idiosyncrasies.

4.3 Predictors

Background data is drawn either from questionnaires administered to students after the test or from school questionnaires administered to principals in each of the selected schools. Here, I describe in detail the predictor variables I used for my estimates.

I follow Soh et al. (2021)'s choice of predictors in selecting the number of books at home as a measure of family backgrounds (instead of the classical PISA composite ESCS index[28], which is not comparable to TIMSS data) and a measure of shortage of instructional material as a proxy for school resources. In line with Anghel et al. (2015), I also include (first generation) immigration status to cover a further dimension of students' socio-economic background. I cannot include student-teacher ratio - as done by Anghel et al. (2015) and Belot and Vandenberghe (2011) - as it is one of the variables that the Gelmini reform aimed to increase: its inclusion may trigger anticipation effects.

These choices are in line with socio-economic background and school resources being major determinants of educational performance, as highlighted in section 3 (although measures of school resources often fail to reach statistical significance). A third factor behind both economic growth and school achievement are institutions. This is why I included a measure of school accountability, reflecting whether schools use assessment scores for comparison with other schools. Unfortunately, this variable is only available in PISA.

PISA questionnaires gather information on student-teacher ratio, shortage of instructional material, and shortage of teachers (this is also available in TIMSS). These would be relevant predictors, because they are carriers of useful information on school resources. Unfortunately, I cannot employ them as predictors due to potential anticipation effects. Indeed, as highlighted in section 2, the Gelmini reform aimed at increasing the student-teacher ratio and cutting on teaching and non-

[27] They exploit the fact that TIMSS 1995 used both variants of the model.

[28] ESCS is a composite measure of socio-economic background, based on parents' occupation, education, and home possessions. Wößmann (2016) found that ESCS enters the education production function only marginally negatively. The literature on international assessments provides evidence that the number of books at home is the most important predictor of student performance, considerably stronger than parental education and occupation (see Hanushek and Wößmann (2011) and Wößmann (2016)).

teaching staff and more generally school resources, in order to reduce spending on education, and thus deficit.

Instead of using them as predictors, I can run my synthetic control again using these three measures of school resources as outcome variables: this is revealing of the mechanisms through which the reform impacted performance. Some researchers argue that measured pupil-teacher ratios (STRATIO) are reasonable approximations of actual class sizes, especially in primary schools. In fact, actual class sizes may be larger than observed STRATIO due to teachers' absences or specialization. Moreover, class sizes would be lower than observed STRATIO in multiple-shift systems (rather than same-time systems).

The "books at home" measure is a multiple choice question such that students can select different ranges (e.g. 0-10; 201-500; more than 500, etc.). Although these ranges change over waves and across studies, it is possible to create a dummy equal to one if the pupil has access to more than 100 books at his place, and this would be comparable over time and across assessments. Immigration status is measured as a dummy for whether a student is a first generation immigrant. Although I did not experiment with alternative measures, it is possible to create a dummy for second generation students. My accountability measure (only available in PISA) is a dummy equal to one when assessments of 15-year-old students are used to compare the school to district or national performance.

Student-teacher ratio is expressed as a positive integer and is comparable across waves since PISA 2000; this measure is based on questions on school enrollment and teaching staff totals. Both shortage of teachers and shortage of instructional material are retrieved from multiple choice questions whose possible values are "Not at all", "Very little", "To some extent", and "A lot" (or expressed with very similar terms). I coded both measures equal to 0 for the first two choices, and 1 otherwise. PISA 2000, PISA 2015, TIMSS 2003 and TIMSS 2007 refer to the shortage of generic teaching staff, while the other waves refer to qualified/experienced teachers in mathematics. Therefore, they are not fully comparable across waves, and this should be kept in mind when interpreting results.

As argued by Ammermüller (2005), transforming variables into dummies makes background variables not only comparable over time within studies, but also across studies.

I imported two additional country-level variables from the World Bank's World Development Indicators, which are used as predictors. The description of these two variables reads:

- *GDP per capita, PPP* (constant 2017 international \$)[29]: GDP per capita

[29] Source: International Comparison Program, World Bank; World Development Indicators database, World Bank; Eurostat-OECD PPP Programme.

based on purchasing power parity (PPP). PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the country plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2017 international dollars.

- *Share of school enrollment in private primary schools* (as a percentage of total primary enrollment)[30]: private enrollment refers to pupils or students enrolled in institutions that are not operated by a public authority but controlled and managed, whether for profit or not, by a private body such as a nongovernmental organization, religious body, special interest group, foundation or business enterprise.

The reason for including them is readily stated. Per capita GDP is often employed in the economics of education literature as a proxy for parents' income. Alternatively, Hanushek and Wößmann (2011) find a correlation of 0.93 between country-level spending in education and per-capita GDP. As highlighted in section 3, there is some evidence that the share of privately owned schools is among the institutional features that impact performance (either directly or through general-equilibrium effects), or at least it interacts with the effect of accountability systems and autonomy measures. I chose the share of private enrollment in primary school rather than lower secondary school both because data is more complete for primary school, and because it delivers a cumulative effect on children education history.

Most of these factors suffer from endogeneity if exploited to estimate an education production function (as in Hanushek and Wößmann (2011)); however, exogeneity is not a requirement for predictors in SCM. Predictors need to be chosen so as to best explain the outcome variable, as long as they were not targeted by the reform itself (generating anticipation effects).

Missing values are a serious concern for researchers analyzing PISA dataset at the micro level. Most of missing values at the student-level are due to non-response[31], while part of missing values at the aggregated level are due to the fact that some questions could not be administered in some countries. Although values are not missing at random, aggregating data at the national level curbs concerns about potential estimation biases, as there is no obvious reason to believe

[30] Source: UNESCO Institute for Statistics (`uis.unesco.org`).

[31] Different sample selectivity can be caused by different enrollment rates, non-response, and exclusion (e.g. for handicapped students).

that countries are differentially affected by missing observations. Country non-participation is discussed in the next subsection.

Tables 1 and 2 contain descriptive statistics of outcome variables and predictors for Italy and a simple average of donor pool countries[32].

Table 1: Descriptive statistics of students in Italy

	TIMSS/99	PISA/00	PISA/03	PISA/06	TIMSS/07	PISA/09	PISA/12	PISA/15
Math score	479.48 (86.53)	- (-)	465.66 (95.69)	461.69 (95.80)	479.63 (76.23)	482.91 (93.04)	485.32 (92.78)	489.73 (93.57)
Reading score	- (-)	487.47 (91.41)	475.66 (100.73)	468.52 (108.75)	- (-)	486.05 (95.88)	489.75 (97.11)	484.76 (93.79)
Books at home	.350 (.48)	.464 (.50)	.413 (.49)	.408 (.49)	.383 (.49)	.394 (.49)	.378 (.48)	.443 (.49)
Immigration status	.030 (.17)	.024 (.15)	.030 (.17)	.050 (.22)	.049 (.22)	.058 (.23)	.071 (.26)	.067 (.25)
Accountability	- (-)	.203 (.40)	.328 (.47)	.197 (.40)	- (-)	.340 (.47)	.651 (.48)	.817 (.32)
Per-capita GDP	41,502 (-)	43,054 (-)	43,781 (-)	44,918 (-)	45,357 (-)	42,075 (-)	41,502 (-)	40,248 (-)
Private ownership	.070 (-)	.066 (-)	.068 (-)	.068 (-)	.070 (-)	.069 (-)	.069 (-)	.064 (-)
STRATIO	- (-)	9.14 (2.29)	10.04 (5.47)	9.19 (2.76)	- (-)	9.35 (2.82)	10.31 (3.37)	10.50 (2.70)
Shortage I.M.	- (-)	.12 (.32)	.24 (.43)	.16 (.37)	- (-)	.18 (.38)	.12 (.32)	.42 (.42)
Shortage teachers	.26 (.44)	.20 (.40)	.20 (.36)	.15 (.38)	.17 (.42)	.22 (.36)	.16 (.40)	.32

Note: Standard deviations in parentheses. TIMSS test scores before rescaling. Per-capita GDP and private ownership are measured at the aggregate level, with no uncertainty.

4.4 Imputation

Abadie et al. (2015) reviews Synthetic Control methods and highlight that they require a *balanced* panel for outcome variables (not necessarily for predictors). Country non-participation is not an issue when restricting observations to PISA waves, as there is no country-year missing observation. It becomes a relevant issue when experimenting with TIMSS data, as only 13 out of 25 countries in my database can be ancored to PISA. I experiment with both linear interpolation and linear regression imputation for missing TIMSS data, after PISA and TIMSS have been merged and TIMSS data have been transformed and anchored to PISA.

Because predictors other than pre-intervention lags of the outcome do not play a relevant role in determining synthetic control weights, imputing missing data for those predictors does not substantially change my results. In my baseline estimate, I imputed missing predictors with linear interpolation.

[32] Rescaled TIMSS test scores would read: Italy 1999 [461.70]; Italy 2007 [461.84]; donor pool avg. 1999 [507.56]; donor pool avg. 2007 [498.85].

Table 2: Descriptive statistics of students in the donor pool (simple average of countries)

	TIMSS/99	PISA/00	PISA/03	PISA/06	TIMSS/07	PISA/09	PISA/12	PISA/15
Math score	511.99	-	502.95	499.24	489.65	497.53	496.32	494.32
	(83.60)	(-)	(93.20)	(92.25)	(82.38)	(92.39)	(92.29)	(89.93)
Reading score	-	498.32	494.77	489.14	-	493.22	497.65	496.67
	(-)	(98.30)	(95.96)	(100.28)	(-)	(95.72)	(97.03)	(98.79)
Books at home	.523	.527	.497	.467	.435	.436	.414	.421
	(.49)	(.49)	(.49)	(.50)	(.49)	(.49)	(.49)	(.48)
Immigration status	.079	.086	.080	.082	.088	.080	.085	.092
	(.27)	(.26)	(.26)	(.26)	(.29)	(.26)	(.26)	(.27)
Accountability	-	.478	.496	.455	-	.578	.673	.699
	(-)	(.40)	(.40)	(.42)	(-)	(.41)	(.41)	(.36)
Per-capita GDP	39,195	40,807	42,600	46,588	48,277	45,935	47,377	49,545
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
Private ownership	.113	.112	.110	.107	.106	.104	.100	.106
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
STRATIO	-	13.09	13.56	12.91	-	12.60	12.79	12.90
	(-)	(3.54)	(3.74)	(3.29)	(-)	(3.42)	(4.44)	(3.85)
Shortage I.M.	-	.20	.27	.26	-	.19	.17	.28
	(-)	(.36)	(.39)	(.40)	(-)	(.36)	(.36)	(.41)
Shortage teachers	.24	.11	.22	.16	.17	.24	.18	.29
	(.36)	(.28)	(.32)	(.37)	(.38)	(.34)	(.39)	

Note: Standard deviations in parentheses. TIMSS test scores before rescaling. TIMSS averages are based on different pools of countries in 1999 and 2007 (due to differences in country participation). Per-capita GDP and private ownership are measured at the aggregate level, with no uncertainty.

5 Methodology

5.1 Synthetic control setting

The main limits in estimating the effects of a change in the institutional features of a country are the followng two: they are impossible to indentify using national data only, because the variation occurs between countries, and most internationally comparable data are cross-sectional (raising concerns of omitted variable bias).

Since Abadie and Gardeazabal (2003) and Abadie et al. (2010), Synthetic Control methods (SCM) have been widely used to study the effects of aggregate interventions and reforms in the fields of economics and social sciences, but also in engineering and medical research. SCM has been described by S. Athey and recent Nobel Prize winner G. Imbens as “arguably the most important innovation in the policy evaluation literature in the last 15 years”. Here, I heavily rely on Abadie (2021) to summarize the main features of Synthetic Controls.

SCM address a key issue that emerged in empirical research during the 1990s: regression analysis is unsuitable to explore the impacts of large interventions at the aggregate level as it requires large samples to carry out inference. Single-unit time series analysis struggles with shocks to the variable of interest aside from the treatment. Finally, the 1990s wave of comparative case studies could not develop credible inference due to the lack of a formalized method for the selection of control units.

Synthetic Control methods provides a systematic way of picking control units that are comparable to the treated, and in doing so they allow to carry out inferential analysis even in the case of one (aggregate) treatment unit only. For these reasons, SCM are often seen as a bridge between large-sample quantitative analyses and more qualitative comparative case studies.

The setting required by SCM includes J observations in the “donor pool” (i.e. the “pool” from which control units are chosen), as well as one ($j = 1$) aggregate treated unit. All units are observed for T_0 pre-intervention periods and T post-intervention periods. We also observe $k \times 1$ vectors of predictors ($\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{J+1}$) for pre-reform periods; these cannot be affected by the reform itself, in order to avoid potential *anticipation effects.* Not only the sample must be a balanced panel, but units in the donor pool should also be unaffected by the same intervention that occured in the treated unit; large idiosyncratic shocks to units in the donor pool should be avoided, too. Moreover, units in the donor pool should be similar to the treated unit to avoid interpolation bias and overfitting[33]. Points in time are not

[33] In order to reduce interpolation bias, Abadie and L’Hour (2021) proposed the use of penalty terms in the minimization problem for units that differ greatly from the trated in terms of predictor variables.

necessarily equidistant.

For $t > T_0$, we want to estimate the effect of the reform, which can ben interpreted as the difference between the observed outcome of interest and the counterfactual outcome (Rubin (1974) we would observe in the absence of the reform (notice that the effect is allowed to change and/or accumulate over time once the reform has been implemented):

$$\tau_{1t} = Y_{1t} - Y_{1t}^N, \quad t > T_0$$

The estimation of the synthetic control boils down to the estimation of a $J \times 1$ vector of weights $\boldsymbol{W} = (w_2, \ldots, w_{J+1})'$ to be assigned to the units in the donor pool so as to simulate a "synthetic" treated unit under the counterfactual state of world where no intervention happened:

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt}$$

In standard applications of the synthetic control, weights are assumed to sum to one and to be nonnegative, preventing extrapolation. This is a further improvement on regression analysis: it can be shown that the regression estimation uses a linear combination of control units, but extrapolation can occur as weights are allowed to go negative, without the reseearcher even noticing it (see Abadie et al. (2015)). Finally, weights are tipically *sparse*, meaning that only a small set of units from the donor pool are assigned positive weights: the researcher can thus check which units have been selected and comment on the credibility of such a selection.

The assumption is that these weights reflect structural parameters that would not vary in the absence of the reform: a similar assumption underlies *difference-in-differences* analyses, where we impose that unobservable differences between treated and control observations remain constant over time. If we are ready to assume that the SC provides a proper counterfactual for our treated unit, then future changes to the predictor variables of the outcome should be deemed as exogenous shocks.

Weights are chosen so that they minimize the difference between pre-intervention values of the predictors for the treated and the synthetic units. Because weights do not depend in any manner on post-reform values of the outcome variables and predictors, SCM does not lend itself to specification searches. The literature established the minimization of the following *norm* as the standard minimization procedure:

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_0\boldsymbol{W}\| = \left(\sum_{h=1}^{k} v_h(X_{h1} - w_2X_{h2} - \cdots - w_{J+1}X_{hJ+1})^2\right)^{1/2}$$

where $\boldsymbol{X}_0 = [\boldsymbol{X}_2, \ldots, \boldsymbol{X}_{J+1}]$ are the predictor values for the units in the donor pool. The constants $v_1, \ldots, v_k$ stand for the relative importance of each predictors variable in predicting Y_{1t}^N. It should be noticed that these predictors usually include pre-reform values of Y_{jt}: the intuition is that matching on pre-reform values of the outcome of interest takes care of other unobserved determinants of performance. In my baseline specification, I include all pre-intervetion outcomes as separate predictors, in line with what is suggested by Abadie et al. (2010). If observed and synthetic Italy are similar over the entire pre-intervention period, then we can expect them to produce similar trajectories even after the reform takes place, unless the reform itself differentially affected the treated unit.

A common way of choosing $\boldsymbol{V}$ is to minimize the mean squared prediction error of the synthetic control relative to Y_{1t}, $\sum_{t \in T_0} (Y_{1t} - w_2(\boldsymbol{V})Y_{2t} - \cdots - w_{J+1}(\boldsymbol{V})Y_{J+1t})^2$, during the pre-intervention period. Alternatives to this methodology include the inverse of the variance of $X_{h1}, \ldots, X_{hJ+1}$ and out-of-sample validation strategies.

As highlighted by Abadie and L'Hour (2021), SCM are conceptually very similar to nearest neighbor matching[34], but they do not impose a fixed number of matches, nor they compute simple averages of control units. It should be noticed that SCM improve on *regression discontinuity designs* in that they do not restrict pre- and post-intervention trends to be linear. For the same reason, it can be considered as a generalization of the standard *difference-in-differences* model (Abadie et al. (2010)), as SCM do not require the effect of unobserved confounders to be constant over time. Still, as in *diff-in-diff* analyses, SCM require the assumption that other factors affecting achievement did not change concurrently in the treated region relative to regions.

The main advantages of SCM are easily said: (i) they prevent extrapolation, (ii) they show the actual discrepancy between observed and synthetic unit in terms of both outcome and predictor variables, (iii) they are less prone to specification searches than other methods (as the estimation of synthetic controls only exploits pre-treatment values[35]), and (iv) they force the researcher to discuss the units contributing to the synthetic control, especially when this is sparse (i.e. few units have positive weights).

[34] Both SCM and nearest neighbor matching assign to matched units positive weights that sum to one.

[35] Synthetic controls can be pre-registered before post-treatment values become available.

5.2 Inference

In Synthetic Control methods, inference is based on permutations: "placebo effects" are estimated by pretending that the intervention occured in each of the other units from the donor pool, in an iterative manner[36]. This means that a new synthetic control is built that matches each unit in the donor pool as best as possible (with different weights each time).

The distribution of placebo effects provides nonparametric *exact* inference[37]. The effect of interest will be considered statistically significant when the estimated effect for the treated unit is "extreme" compared to the placebo effects. This is especially true when the effect of the intervention is large relative to other determinants of the output.

A major issue with such a naïve inferential method is that some of the units in the donor pool may not be matched as well as the treated unit is. One way of dealing with this issue is to discard units for whom approximation is deemed poor compared to the unit of interest: in practice, we measure the pre-intervention root mean squared prediction error (RMSPE), which is the preferred measure of "goodness of fit" in SC analysis:

$$R_j(0, T_0) = \left(\frac{1}{T_0 + 1} \sum_{t=0}^{T_0} (Y_{jt} - \hat{Y}_{jt}^N)^2 \right)^{1/2}$$

Then, we exclude units whose post-reform RMSPE is greater than a multiple of our treated unit's RMSPE; typical thresholds are 2, 5 and 20. An alternative to this method is to compute the ratios of post- to pre-intervention RMSPE for all units, and compare them. Eventually, the p-value is computed as the ratio of the number of units for which the post-intervention effect is at least as large as the effect for the treated unit (treated unit included), to the number of units considered (treated unit plus donor pool counries), once we have taken care of pre-intervention fit appropriately.

In most cases, the mechanism by which treatment is assigned to a unit is not random; moreover, such an assignment mechanism cannot be rationalized (what was the probability the Gelmini reform would be implemented in Germany?). By formally defining a procedure to create a synthetic comparison unit, SCM allow the estimation of placebo treatments, and thus quantitative inference. A final remark should be made: contrary to sampling-based statistical inference, permutation in-

[36] These are technically referred to as "in-space placebos". I do not delve into "in-time placebos" as they cannot be applied to my analysis due to the lack of sufficient data. See Abadie (2021).

[37] This mode of inference is *exact* because we can always compute the distribution of the estimated effect for placebo units, independent of the number of control units and points in time, and regardless of the use of aggregate or micro-level data.

ference does not define a sampling mechanism through which units are sampled, and often the sample concide with the population; variation in the test statistic derives from the assignment mechanism only, conditioned on the available sample. When using aggregate data, traditional standard errors should return a p-value of exactly 0, as there is no uncertainty about the value at, say, the national level. However, the point is that permutation inference addresses the remaining concern that the synthetic control might not be the "perfect" control and there is uncertainty about the true counterfactual state of the world (i.e. the uncertainty about the ability of the synthetic control to reproduce the movement of the counterfactual unit in the absence of treatment). This source of uncertainty is typical of comparative case studies such as SCM. The use of micro-data, as it is the case in this work, simply increases the total degree of uncertainty, but permutation tests remain the preferred mode of inference. Using placebos allows the researcher to explore potential hidden biases by looking at what happens to other countries that - at least in principle - should not be exposed to the treatment.

Much of the statistical theory of causal inference traces back to the work by R. Fisher, who addressed the issue of heterogeneity with randomization. This approach to causal inference was opposed to the one proposed by J. S. Mill, who suggested to eliminate as much heterogeneity as possible in experimental trials and compare "two cups that are alike". Rosenbaum (2005) notices that, often in observational studies, randomization is not possible; in most SC analyses, the assignment mechanism is indeed not random. He proceeds to prove that, when randomization is not feasible, reducing heterogeneity among observed units provides larger inferential advantages than simply increasing the sample size, as it reduces both sampling variability and sensitivity to unobserved bias (selection into treatment), while increasing samples only addresses sampling variability. Carefully selecting control/matching units - as SCM do - is a way of reducing heterogeneity in the spirit of Rosenbaum (2005).

5.3 A few examples

To name just a few applications of SCM, Abadie and Gardeazabal (2003) first applied SC to the effect of 1970s outbreak of terrorist attacks in the Basque Country on economic output; here, inferential analysis was still in an embryonal phase, with only one unit from the donor pool serving as "placebo study". Abadie et al. (2010) look at the 1988 tobacco control program implemented in California, exploiting yearly tobacco sales data for the 1970-2000 period; the other US states act as donor pool. Pinotti (2015) studies the 1980s outbreak of mafia organizations in Apulia and Basilicata with 17 other Italian regions serving as donor pool, and using decade-long data on GDP and predictor variables. Bohn et al. (2014) analyse the impact of a 2007 Arizona act tightening work eligibility for new immigrants, exploiting nine years of Current Population Survey data as pre-intervention period and 46 other US states as donor pool. Abadie et al. (2015) exploit 1960-2003 GDP time series for 17 OECD countries to estimate the effect of German Reunification on West German economic outcome. Acemoglu et al. (2016) use time series data on a large number of US stocks to gauge the impact of the nomination of T. Geithner as Treasury Secretary (2008) on abnormal returns for firms that were personally connected to his figure. Peri and Yasenov (2015) re-consider the Mariel Boatlift and exploit SCM to improve on the 1990 analysis by David Card: they compare the labor market outcomes for Miami with those of 43 other US metropolitan areas using six pre-intervention observations from the Current Population Survey.

5.4 Empirical framework

In this work, I apply Synthetic Control methods to the PISA database, in order to estimates the effect of the Gelmini school reform on Italian students' performance in mathematics and reading. Before moving on to the results, I must discuss why I employed SCM and whether this methodology can accomodate my research question, given the available datasets.

A few reasons stand out for applying synthetic controls to approach my research question. First and foremost, they allow me to separate the impact of the Gelmini reform from that of long-term trends[38] and common shocks (e.g. the 2008 crisis). Moreover, although OECD and IEA put effort into ensuring score comparability over waves through "linking items", questions were different across different waves: studying the difference between observed and synthetic score curbs this issue. Finally, although countries implementing large reforms at the national level were excluded from my donor pool, at the micro level, schools and teachers may still respond to previous waves of international assessments (possibly encouraged by central administrations), and try to improve pupils' performance in standardized tests (e.g. through *teaching to the test*). If we are ready to assume that such reaction was similar for countries with close enough scores, then SC can address these concerns.

A couple of technical notes should be added to this discussion. Firstly, by aggregating data at the country level, I circumvent potential selection problems (e.g. students attending private schools along observable or unobservable dimensions), I am able to uncover general equilibrium effects (e.g. the competition effect of private schools), and more generally I make spillovers more implausible, lending

[38] Over the last few decades, most countries around the world have experienced a growth in average school attainment and a reduction in its dispersion.

grater credibility to the *Stable unit treatment value assumption* (SUTVA). Secondly, although the Gelmini reform needs not to be as good as randomly assigned to the treated unit, the empirical strategy requires that the enactment of the reform is exogenous to trends in the outcome variable. Moreover, the 2008 economic crisis should not drive the results: although this cannot be ruled out in the data, there is no obvious reason why I should doubt that the economic crisis differentially impacted the performance of Italian students. It is true that Italy was hit harshly by the crisis (and the subsequent sovereign debt crisis), but this only means that my estimates are a lower bound for the true effects of the reform.

The modest volatility of the outcome variable (see below), the availability of a number of comparable units, the arguable lack of anticipation and spillover effects, and the availability of a proper time horizon[39] are all features of the setting and the aggregate data that create a suitable framework for the application of Synthetic Control methods. One should also notice that the feasibility of my empirical strategy relies on the SCM accommodating microdata. For instance, Peri and Yasenov (2015) exploit the Current Population Survey (CPS) to re-estimate the effects on the Miami labor market of the Mariel Boatlift. Notice that, although they rely on a few more pre-intervention periods, the CPS was administered to just about 40 individuals per unit-year, while PISA is administered to more than 2000 students per country in each wave.

One major drawback is shared by studies that apply SCM to international assessments (and a few other studies, such as Peri and Yasenov (2015)), thati is, the scarcity of pre-intervention period data. Indeed, their credibility relies on their ability to closely track the pre-reform trends in the outcome variable (moreover, there are technical details explained in section 3.3 of Abadie et al. (2010)). Unfortunately, PISA data administered only two comparable pre-intervention waves for mathematics, and three for reading. Scarsity of pre-intervention periods has a few consequences on synthetic control estimation and inference, which I discuss in the following subsection.

5.5 Addressing concerns of data availability

Johnson (2013) finds that the application of SC to international assessment scores has a few idiosyncrasies that must be discussed here. First and foremost, covariate predictors tend to play nearly no role in determining SC weights. Although this is in contrast with other applications of SCM to political science and economics, this is an expected result given the high intercorrelations of achievement measures, so that

[39] In other words, there are enough post-intervention periods to see the medium-term effects of the reform.

covariate predictors cannot really say much more than what is already explained by the lags of the dependent variable. In fact, this feature of educational assessment studies imply that an inappropriate choice of predictors should not exaggerate the unobserved matching bias, and it is thus an advantage of this specific setting[40]. The bad news is the synthetic control will hardly resemble the treated unit in terms of covariate predictors (no goodness of fit), making results less interpretable, and thus less reliable.

A second point to be discussed is that, for some units, the SC is able to perfectly match pre-intervention values of the predictors: this occurs in international assessment settings because few pre-intervention periods are usually available and the predictors (at least those that receive a positive weight) of some units happen to belong to the *convex hull* of donor pool units' predictors. For other units, this condition is not fullfilled, and the unit's values are only aproximated. The convex hull condition is a technical one in SCM, and one that is often not fullfilled in empirical applications[41]. In fact, the validity of such condition has some drawback that I discuss here.

When a unit "belongs" to the convex hull of its control units, its synthetic control may not be unique nor sparse (a synthetic control is *sparse* when few control units receive positive weights). In such a setting, increasing the number of pre-intervention points is one way to deal with this issue. In particular, Johnson (2013) finds that adding additional pre-treatment events reduces the probability of obtaining close or perfect matches, simultaneously reducing the required effect sizes for statistical significance.

Alternatively, sparsity can be increased by imposing a bound on the number of nonzero weights. In particular, a recent work by Abadie and L'Hour (2021) introduced a variant of SC that penalizes discrepancies between the unit and control regions' predictor values[42]: as a result, interpolation bias is curbed and it can be proved that this variant always generates unique and sparse synthetic controls. In robustness, I implemented this new methodology to impose sparsity to my SC[43]. As an additional measure, I merged PISA and TIMSS databases to increase the probability of obtaining sparse controls.

Lack of sparsity makes interpretability of results more complex. Moreover, overfitting may arise when the characteristics of the treated unit are artificially

[40] Moreover, it means that the volatility of the outcome is modest, so that a substantial shift in achievement caused by reforms should be easily distinguishable from other shocks.

[41] Failure to meet the convex hull condition does not invalidate the results of SCM.

[42] In settings with close/perfect matches, the researcher faces a trade-off between minimizing the distance between the treated unit and its SC, and minimizing the distance between the treated unit and each unit composing the SC. The penalized SC introduces the latter term in the minimization problem.

[43] Check https://github.com/jeremylhour/pensynth for Abadie and L'Hour (2021)'s code.

matched by combining the values of a large sample of unaffected units. Not surprisingly, most authoritative SCM studies exploit data with more than 20 pre-intervention periods. My work shares the scarcity of pre-reform data with the literature on SC applied to international assessment data (as well as with other papers from different branches of economics and political science, such as Peri and Yasenov (2015)). This is why I devoted much effort to robustness analysis.

I use the R package `tidysynth` to run my analysis. This is based on the `Synth` package described in Abadie et al. (2011).

6 Results

6.1 Baseline estimates

In this section, I present the baseline results and comment on them. In section 7, I show the robustness analysis. However, before diving into the results, I discuss my priors on the effects of the reform.

If one had to believe the reasons of the students and teachers' protests following the enactment of the Gelmini reform - and, more generally, those of the movement that goes under the name of "Onda anomala" -, negative effects on pupils' achievement would naturally be expected. Remember from section 2 that, among the others, the Gelmini reform increased minimum and maximum class sizes in elementary and secondary school and decreased weekly instruction hours in middle school, with the aim to drastically reduce teaching staff; it also reduced non-teaching staff and embarked on a series of general spending cuts. Battistin et al. (2015) provide some evidence that the reform weakened the performance of elementary students in national Invalsi tests by 4%, but the reform did not achieve its goals in terms of student-teacher ratio and class size: the impact of the reform on these variables was indeed modest.

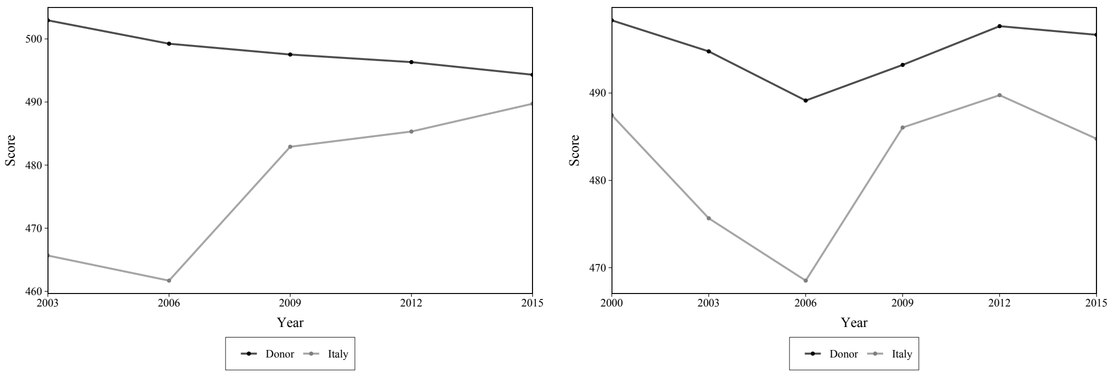
In fact, the intention of the reform was to promote a re-organization of the Italian school system that would positively impact students' performance (see art. 1, Law 133/2008). Mariastella Gelmini stated that the reform only aimed at cutting wasted and unnecessary resources (LaRepubblica (2009)). One may also argue that the reform only affected class sizes at the margins (minimum and maximum class sizes were increased from 15 to 18, and from 25 to 27, respectively), but it did not reach most of the classes, which lie in the middle of the distribution. Moreover, the reduction in teaching staff hit the less qualified teachers that ended up at the lower end of recruitment rankings. Finally, as part of the reform, Gelmini introduced the Invalsi test as an additional external examination in final middle school exams (low stakes Invalsi tests were also introduced in elementary school, but this may impact pupils' achievement only in later PISA waves). Because Invalsi and PISA (but also TIMSS) share numerous common features, one may expect that 8th grade students' preparing for math and reading Invalsi tests (since s.y. 2007/8) improved their performance in Invalsi-like tests, such as PISA (since the 2009 wave). In line with this argument, results may differ from those by Battistin et al. (2015) not only because I use PISA database instead of Invalsi results, but also because the introduction of Invalsi itself as part of middle school exit exam may have impacted performance in PISA.

Fortunately, as explained in section 5, I can separate the effect of Invalsi from those of other provisions. Indeed, the introduction of the Invalsi test is likely to boost performance in standardized tests as soon as it is implemented[44]. On the contrary, changes in class size and instructional time only affected newly formed classes since s.y. 2009/10. Given that PISA tests 15 year old students, the impact of these changes should only be seen since PISA 2015 wave. It should be noticed that this effect is *cumulated* over three years of middle school (lower secondary), and that no effect should be observed until PISA 2012 wave (included). I can also separate the effect produced by the reform of high schools: the provisions included a drastic cut in instruction time, and a re-organization of high school tracks; however, the reform was only implemented since s.y. 2010/11, meaning that the effect should be seen since PISA 2012 wave. Battistin et al. (2015) had to assume that changes in instruction time and class size that directly affected newly formed classes did not indirectly impact other classes in the school through re-organization choices by school principals. I do not need to make such an assumption, as the PISA 2009 wave predates the implementation of all major changes except the introduction of Invalsi tests in middle school.

A first, expected results is that a simple average of all countries in the donor pool cannot approximate Italian achievement. Figure 1 shows that this is the case for math and reading performance in PISA, although reading scores present similar trends, while math trends are by no means comparable (this is in line with baseline results, below). Still, a weighted average of donor pool units is likely to deal closer approximations: Synthetic Control methods provide a data driven approach to choose such weights.

The presentation of my results begins with synthetic controls for mathematics

Figure 1: Italy and donor pool (average) scores



(a) Mathematics

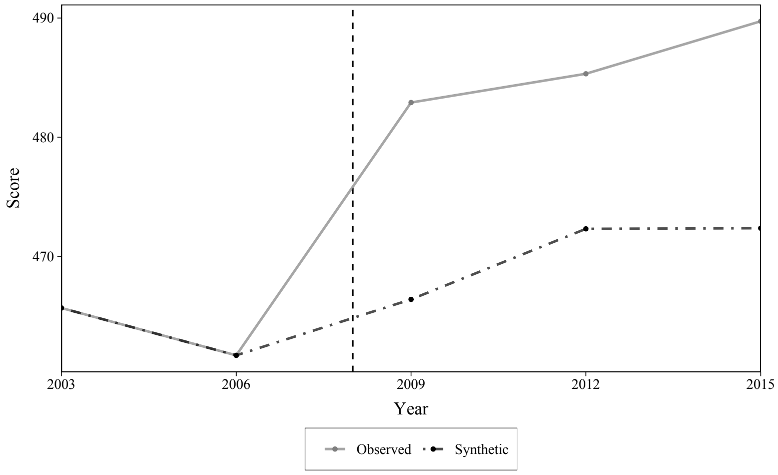
(b) Reading

[44] Invalsi had been administered to samples of students even before s.y. 2007/8, meaning that students and teachers knew what to expect from the new test, and could prepare for it since its first national administration.

and reading score, based on PISA data only. These are my baseline results, while additional insights can be gained from the inspection of results based on the merged PISA and TIMSS databases (see below). One should remember that math score matching is based on two pre-intervention periods, while reading score matching is based on three. This is because PISA 2000 wave did not link its math scores to subsequent PISA waves.

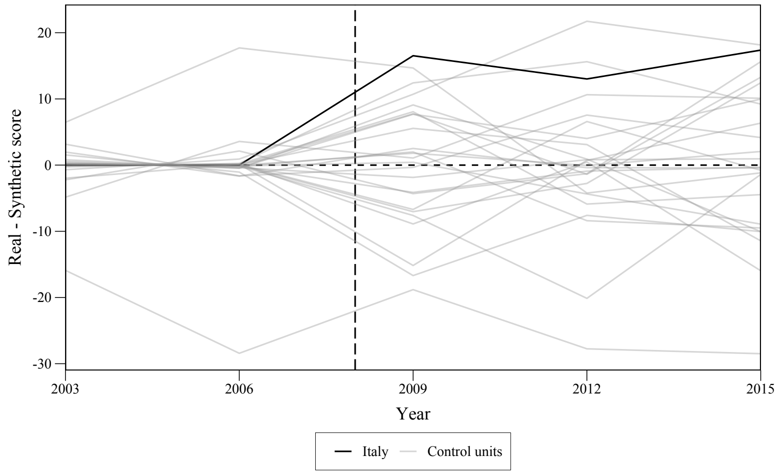
Figure 2 plots the observed math trend for Italy, together with the synthetic trend. Figure 3 plots the difference between the observed and synthetic math trends for Italy and all donor pool units[45]. The second graph illustrates the inferential strategy based on permutations.

Figure 2: PISA math synthetic control



Note: vertical dashed line denotes the time of the intervention.

Figure 3: PISA math permutation inference

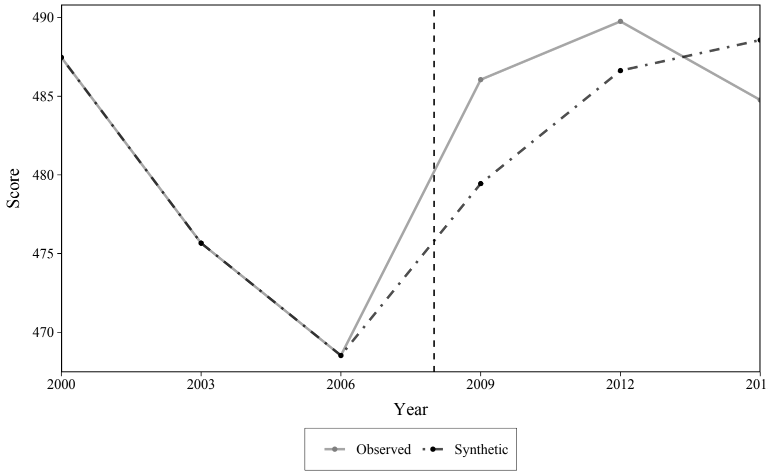


Note: vertical dashed line denotes the time of the intervention.

[45] Notice that, in the case of Italy, the plotted line is the difference between the two trends (observed and synthetic) in Figure 2.

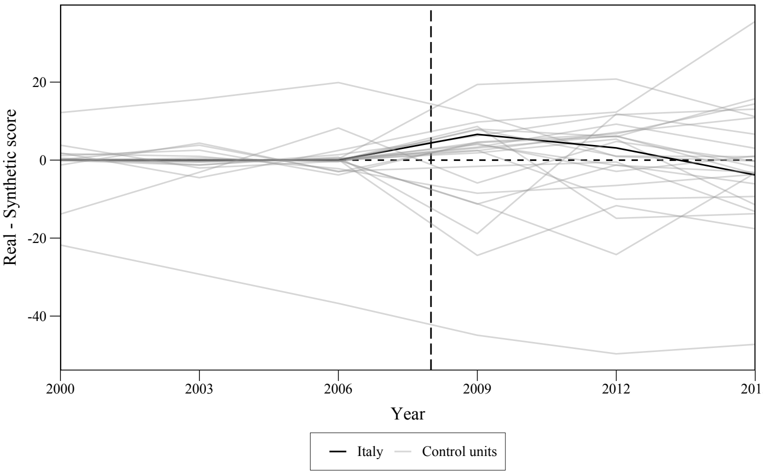
Similarly, Figure 4 plots the observed reading trend for Italy, together with the synthetic trend. Figure 5 plots the difference between the observed and synthetic reading trends for Italy and all donor pool units.

Figure 4: PISA reading synthetic control



Note: vertical dashed line denotes the time of the intervention.

Figure 5: PISA reading permutation inference



Note: vertical dashed line denotes the time of the intervention.

Both math and reading synthetic controls are based on the following predictors: all pre-intervention lags of the outcome, the number of books at home, immigration status, my accountability index based on PISA questionnaires, per-capita GDP, and the share of private enrollment.

A few technical aspects should be dealt with before discussing the results (see section 5). Table 3 presents predictors' weights for mathematics and reading scores. Table 4 compares pre-intervention values of predictors, which is revealing of the goodness of fit. I do not include a table with unit's weights due to lack of sparsity: the weights assigned to donor pool units will be discussed in text.

Table 3: Predictors' weights

	Mathematics (1)	Reading (2)
PISA 2000 score	-	.35
PISA 2003 score	.48	.30
PISA 2006 score	.47	.34
Books at home	0	0
Immigration status	0	0
Accountability	0	0
GDP per capita	.05	0
Private ownership	0	0

Table 4: Observed versus synthetic Italy

	Mathematics			Reading		
	Observed (1)	Synthetic (2)	Donor sample (3)	Observed (4)	Synthetic (5)	Donor sample (6)
PISA 2000 score	-	-	-	487.47	487.46	498.32
PISA 2003 score	465.66	465.66	502.95	475.66	475.66	494.77
PISA 2006 score	461.69	461.69	499.24	468.52	468.52	489.14
Books at home	41.0	46.1	48.2	42.8	48.3	49.7
Immigration status	4.0	8.5	8.1	3.5	6.7	8.3
Accountability	26.2	42.5	47.6	24.2	56.5	47.7
GDP per capita	44,350	44,350	44,594	43,918	34,872	43,331
Private ownership	6.8	7.9	10.8	6.7	11.2	11.0

Note: Variables with [0, 1] range multiplied by 100. All variables, except PISA lags, are averaged at the country level for the pre-treatment period.

First of all, one should notice that covariate predictors play nearly no role in determining SC weights: this result is explained by the high intercorrelations between consecutive achievement scores. This is both good and bad news: the good news being that an inappropriate choice of predictors should not increase unobserved matching bias (unsurprisingly, removing one covariate predictor at a time deals only an impreceptible change in synthetic control estimation); the bad news is the synthetic control does not resemble true Italy except for pre-intervention values of the outcome (math and reading scores). Indeed, as Table 4 shows, my synthetic control only slightly improves on the (simple average of) donor sample in matching Italy's covariate predictors' values.

Secondly, for some units (Italy included), the SC is able to perfectly match pre-intervention values of predictors. This occurs because predictors' values for those units belong to the convex hull of control units' predictors. Once again, there are good and bad news: while perfect matching is certainly desirable, it prevents both sparsity and uniqueness, possibly generating overfitted estimates. One should also notice that, when running placebo treatments, not all countries fall in the convex hull of their donor pool units, meaning that some of them are imperfectly approximated, and a few of them have a poor match.

Small weights assigned to covariate predictors and lack of sparsity curb the interpretability of the results, which is one of the strengths of Synthetic Controls, compared to regression and traditional case studies. In other words, the choice of Italy's synthetic control is a "black box". However, there is something I can do about it, either to increase sparsity, or at least to show that the estimates are robust:

- in robustness analysis, I divide the pre-intervention period in a *matching* and a *validation* period, and show that, at least for math scores, my synthetic control is still able to closely approximate Italy's score trends during the validation period; moreover, I can apply small changes to the donor pool units and predictors;
- Johnson (2013) finds that the inclusion of additional pre-intervention periods reduces the probability of obtaining perfect matches (this is also in line with the discussion in Abadie (2021)): therefore, one thing I can do in robustness analysis is to include more pre-reform events by merging TIMSS data with PISA baseline database;
- finally, I estimated the penalized synthetic conrol recently introduced by Abadie and L'Hour (2021)[46].

In sum, these estimates should be taken with caution. In particular, as highlighted above, the lack of sparsity means that - for both subjects - all donor pool units contribute with positive weights (all greater than .005) to Italy's synthetic control. For mathematics score, Bulgaria, Israel and Luxembourgh contribute with weigths greater than .1, while Russia, the Netherlands, Bulgaria and England contribute mostly to matching Italian reading scores. In this situation, justifying the choice of countries with bigger weights is hopeless, as they are simply chosen so as to match Italian performance with perfection. In robustness analysis, I show how this issue can be at least partially addressed. Still, the ability of the synthetic control to spot post-reform turning points is encouraging. Finally, reverse causality is addressed by the design of my identification strategy, as it is arguable that countries with similar score may have responded in a similar manner to the announcement of their

[46] Although there is no package to conduct this frontrunning methodology, I was able to adapt Abadie and L'Hour (2021)'s `github` code to my synthetic control (for mathematics scores).

performance in previous waves: endogenous response to the lags of the outcome is at least partially controlled by synthetic scores.

6.2 Discussion and inference

It is reassuring that the synthetic control can track the change in trend direction occured since PISA 2009, both for mathematics and reading. This means that, although the choice of units' weights is not as clear as in other economic applications of SCM (though more interpretable than standard regression estimations), still my synthetic control is able to spot turning points that occur after the matching period ends (remember that no post-intervention data is used in the computation of weights).

Results differ substantially by subject. My synthetic control is able to track the dynamic of reading scores quite precisely even after 2008. This means that post-reform changes in performance can be well explained using only pre-intervention data that cannot be influenced by the reform (by construction). By inspecting Figure 4, one may suggest that Invalsi tests somehow boosted reading performance in 2009 and the reform's cuts curbed it slightly in subsequent years, but the effect is not statistically significant and nothing can really be said about it. Indeed, the lack of substantial policy impacts is confirmed by looking at the observed/synthetic score difference for Italy and donor pool units (Figure 5): the difference is close to zero for the entire period and is by no means "extreme" compared to placebo treatment effects.

On the contrary, mathematics performance jumps by 15.64 test-score points compared to the synthetic control. The synthetic control is not able to explain the post-reform breakpoint; moreover, this jump is also at the upper tail of the distribution of placebo treatment effects. In order to gauge the magnitude of this estimate, one should consider that the Italian (within-country) standard deviation of math scores was 93.04 in PISA 2009 and 93.57 in PISA 2015: this means that - if we are ready to consider the observed increase in math performance as a result of Gelmini's intervention - the reform increased Italian students' performance by roughly 16.7% of a (within-country) standard deviation, which is economically significant[47]. This effect is comparable to the one estimated by Mandel and Süssmuth (2011) for an increase in instruction time by one hour per week, cumulated over nine years of instruction. In 2009, Italy scored 33th in mathematics among participating countries: if Italy had scored 15.64 less than it did, it would have ranked 36th[48]. For further comparison, Wößmann (2016) sets, as a rule of thumb, that

[47] The same figure can also be expressed as a proportion of an international standard deviation, which is normed to 100.

[48] Unsurprisingly, it ranked 36th in PISA 2006.

one-year learning gains on most international tests equal between 1/4 and 1/3 of a standard deviation (i.e. 25-30 test-score points in PISA). Whatever the cause of Italy's boost in math performance, the jump between PISA 2006 and PISA 2009 was economically significant.

What we learn from these results is that the reform's cuts to school resources and staff did not cause an appreaciable, negative impact on pupils' PISA achievement. Conversely, it seems to have boosted mathematics performance since 2009. It is arguable - from what was said in section 2 - that the one change that could cause such an early response to the reform was the introduction of Invalsi tests in middle school exit exams. Indeed, all the other interventions were implemented after the administration of the PISA 2009 wave. My take on the results is that teachers' and students' preparation for the newly introduced Invalsi tests may have improved pupils' performance in Invalsi-like standardized tests such as PISA. Other interventions that targeted the efficiency of the Italian education system must have taken more time to be reflected on international tests performance.

The cumulative effect (over three years of middle school) of the reform's cuts should be reflected on PISA 2015 scores, but Figure 3 does not show any detectable difference between 2009 and 2015 observed/synthetic differences in scores. Moreover, if one believes the Invalsi argument, then she should also argue why this is reflected on math scores and not so much on reading scores. One explanation may be that training for reading and comprehension questions is not as effective as training for logic and math multiple answers exercises.

One may also ask whether math results are statistically significant at standard thresholds. In fact, inference is complicated by units' predictors falling into the convex hull of donor pool units' values. Here, I explain why.

Section 5 discussed how to carry out inference in SC applications. The most naïve way of gauging statistical significance is to simply compare post-intervention discrepancy between observed and synthetic scores; in particular, such discrepancy is often computed as the root mean squared error (RMSE). Fitzpatrick (2008) employs this procedure to make inference on NAEP data. In my case, this simple approach is unfair to those countries that have been closely approximated in the pre-reform period, as there are at least two countries (Finland and Bulgaria) that were poorly matched in the pre-Gelmini period and may be "extreme" even after 2008 just because there are no countries in the donor pool that can provide a proper control.

One way Abadie (2021) proposes to deal with poorly matched placebo units is to compute the ratio of post to pre-reform RMSE and gauge statistical significance based on this new ranking. When applying this methodology, Italy has by far the most extreme treatment effect (implying a p-value equal to .04, significant at the 5%

level): however, this is not the result of Italy possessing the largest post-treatment RMSE, but rather the smallest pre-reform one. Indeed, those countries that were perfectly matched in the matching window have pre-reform RMSEs that tend to zero, meaning that the ratio of post- to pre-reform RMSE explode; it is only by chance that Italy is the best match (it has more decimal zeroes).

One last way of making inference, suggested by Abadie (2021), consists in merging the first two methods: one should compare post-intervention RMSEs, but only within a group of units whose RMSE is under a certain arbitrary threshold. For instance, Soh et al. (2021) keep only countries with pre-reform RMSPE smaller or equal to 1. As there is no clear ("less arbitrary") threshold, I start from the ten best matched countries and keep adding one placebo unit at a time, each time evaluating statistical significance with the first method. Doing so, the *p*-value for math scores oscillates between .083 and .154 (Switzerland is the only country with both larger post-reform RMSPE and good pre-reform fit).

In sum, there is modest evidence that the (positive) treatment effect for math scores is statistically significant, which confirms the visual inspection of Figure 3. At the very least, the Gelmini reform does not seem to have negatively affected pupils' performance in international assessments.

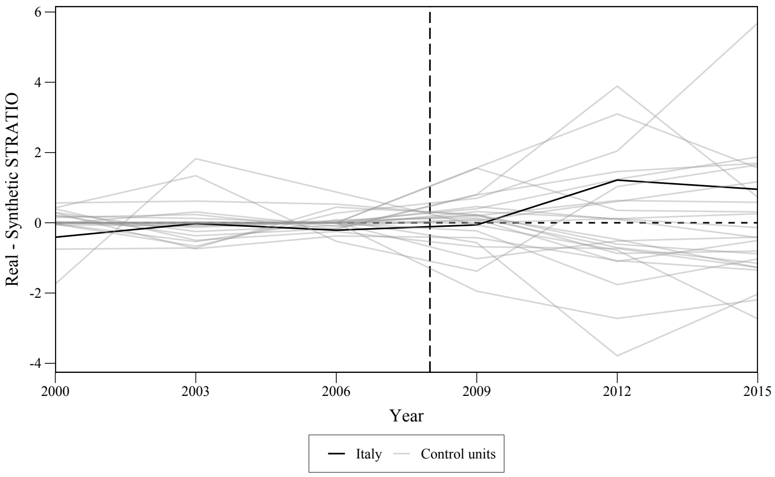
6.3 Mechanisms

In order to enrich my analysis and try to give an explanation for the results I presented in previous subsections, I re-run my synthetic control using student-teacher ratios, shortage of instructional materials, and shortage of teachers as outcome variables (this is also available in TIMSS). These variables measure key school resources items that were targeted by the Gelmini reform. Scores in tests, per-capita GDP and private ownership were used as predictors, along with lags of the outcome variable.

Figures 6 to 8 plot the difference between observed and synthetic Italy for each of the three variables, together with permutation runs. Although the reform was implemented after the administration of the PISA 2009 wave, I decided to match resource variables until 2006 only, in order to exclude potential response bias due to reform announcement. In this regard, it is at least reassuring that student-teacher ratio increases only since PISA 2012.

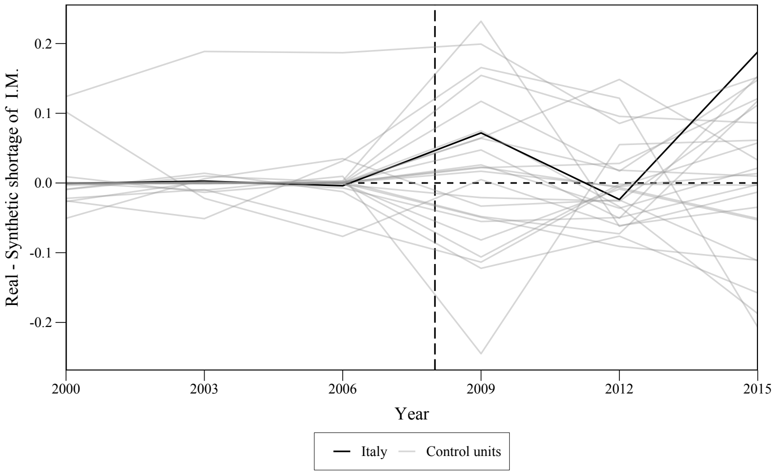
The post-reform trends in all three variables point in the expected direction. In particular, the student-teacher ratio increases by one percentage point since PISA 2012, which is precisely at the target set by Law 133/08. The trend for shortage of instructional materials is not clear, altough it spikes in 2015. It must be noticed that instructional materials proxy for family resources, along with school resources;

Figure 6: Synthetic control for STRATIO (PISA only)



Note: vertical dashed line denotes the time of the intervention.

Figure 7: Synthetic control for Shortage of instructional material (PISA only)



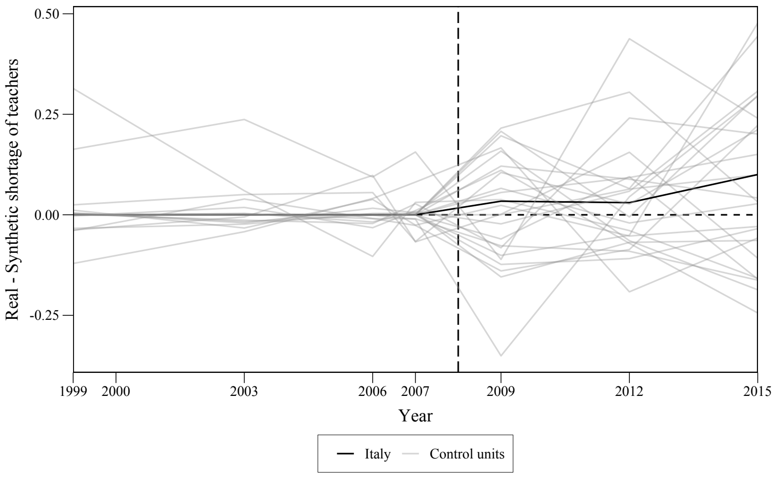
Note: vertical dashed line denotes the time of the intervention.

if this is the case, then the 2015 jump might be explained by Italian families being more harshly hit by the sovereign debt crisis and its consequences on the Italian economy. Finally, the shortage of teachers' measure shows only a slight upwards trend.

These three graphs provide only a modest evidence that the reform reduced Italian schools' resources, and none of these results is statistically significant. This is in line with Battistin et al. (2015), who found that the Gelmini reform did not achieve its goals in terms of student-teacher ratio and class size, at least in elementary schools.

Section 3 has already shown that the literature does not agree on the effect of school resources on pupils' performance. The evidence from this subsection pro-

Figure 8: Synthetic control for Shortage of teachers



Note: vertical dashed line denotes the time of the intervention. PISA 2000, PISA 2015, TIMSS 2003 and TIMSS 2007 refer to the shortage of generic teaching staff, while the other waves to qualified teachers in mathematics. Therefore, they are not fully comparable across waves, and this should be kept in mind when interpreting results.

vides an additional explaination for the lack of observed reductions in international tests performance. On the contrary, the introduction of Invalsi tests and other measures to support the efficiency of the education system may more than compensate potential negative impacts of these modest reductions in schools' resources.

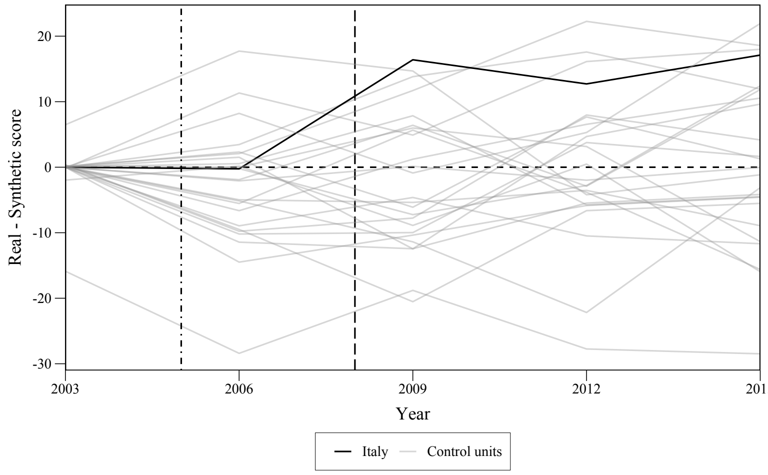
7 Robustness

In previous subsections, I highlighted the issues related to the scarcity of pre-reform periods. In this section, I try to (i) show that my estimates are reliable regardless of the shortage of data, and (ii) address the issue directly, in two distinct ways.

First of all, I already showed that my synthetic control is able to nearly perfectly track the dynamic of reading scores, even though a turning point occurs after the matching period.

Another way to show the validity of my estimates is to cut the pre-reform period in a *training period* and a *validation period*: in other words, I estimate SC weights on math scores based on PISA 2003 only, and inspect the synthetic control's ability to track Italian performance in 2006 (i.e. before the reform was enacted). Figure 9 plots the difference between the observed and synthetic math trends for Italy and all donor pool units. Surprisingly, Italy is perfectly matched in 2006, and the post-reform dynamic is nearly identical to that in Figure 3. Even though the scarcity of pre-reform periods impairs the interpretability of my synthetic control, the algorithm seems to work well in approximating the dynamic of PISA scores (at least for Italy): as suggested by Johnson (2013), this may be the result of high intercorrelations between different waves of international assessments.

Figure 9: PISA math SC with validation period



Note: vertical dot-dashed line denotes the beginning of the validation period, while vertical dashed line denotes the time of the intervention.

One obtains similar results when matching reading performance based on PISA 2000 and PISA 2003 only, leaving PISA 2006 as the validation period (results available from the author): the synthetic control matches PISA 2006 quite well and the dynamic is similar to Figure 5.

Additional confidence on the estimates comes from a *leave-one-out* re-analysis. Indeed, if excluding a unit from the donor pool has a large impact on treatment effects without a significant change in pre-reform matching, one may wonder whether the estimated effect is caused by the impact of idiosyncratic interventions in the excluded control country. I proceed by re-running my mathematics synthetic control excluding each of the three countries that receive weights greater than .1 in baseline estimates (Bulgaria, Luxembourgh, Israel). Reassuringly, the plot of observed versus synthetic scores only slightly changes, and in all three cases, the discrepancy for Italy is more "extreme" compared to placebos runs, relative to baseline estimates (results available from the author).

Finally, I re-run my synthetic control again, removing one covariate predictor at a time. As one may expect from the small weights assigned to all of them in baseline estimates, qualitative and quantitative results are unaffected (results available from the author).

So far, I showed that, even though perfect matching curbs the interpretability of weights assignment, my estimates are robust to changes in the matching period, in predictors, and in donor pool units. Before concluding, I experiment with two solutions to the issue of the lack of sparsity. As stated above, perfect matchings (and thus non-sparse solutions) are more likely to occur with few pre-intervention periods. Therefore, I try to address it by including TIMSS mathematics data to increase pre-reform periods from two to four (the matching procedure is described in section 4). Unfortunately, TIMSS data is only available for 13 out of 25 countries[49], meaning that imputation was necessary for the remaining ones (SCM require balanced panels). Moreover, the procedure relies on the assumption that the merge produces a consistent transformation of scores, which cannot be proven. These are the reasons why I preferred the PISA-only estimates as baseline results.

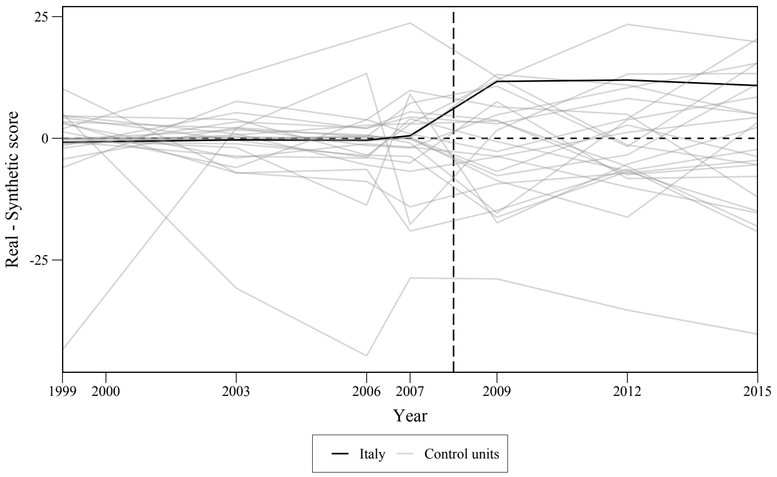
Figure 10 shows observed/synthetic differences in scores for Italy and placebo units. The results are much in line with those from PISA data only: we observe a consistent and lasting, positive effect on performance (notice that post-intervention data is from PISA data only, as TIMSS waves are only used for matching).

What we gain from adding two pre-intervention points is to make perfect matching less likely. Indeed, Italy's synthetic control is now sparse, with only four countries receiving positive weights: Norway, Israel, Bulgaria and the USA (.45, .23, .19, .11, respectively). Not surprisingly, these countries received large weights even in the non-sparse (PISA-only) estimate. Particularly noteworthy are the positive weights assigned to per-capita GDP and share of private enrollment. One explanation is that, with longer pre-intervention data, pre-reform trends are now more visible and covariate predictors become more relevant in explaining long-run dynamics.

There is an alternative solution to the lack of sparsity that does not require

[49] Moreover, the accountability measure is not available.

Figure 10: Merged PISA and TIMSS mathematics scores



Note: vertical dashed line denotes the time of the intervention.

TIMSS data (and, thus, rescaling bias): Abadie and L'Hour (2021) have recently introduced a penalized Synthetic Control method that includes in the minimization function an additional term for the discrepancy between the unit and control regions' predictor values[50]. This term forces the synthetic control to find a unique and sparse solution.

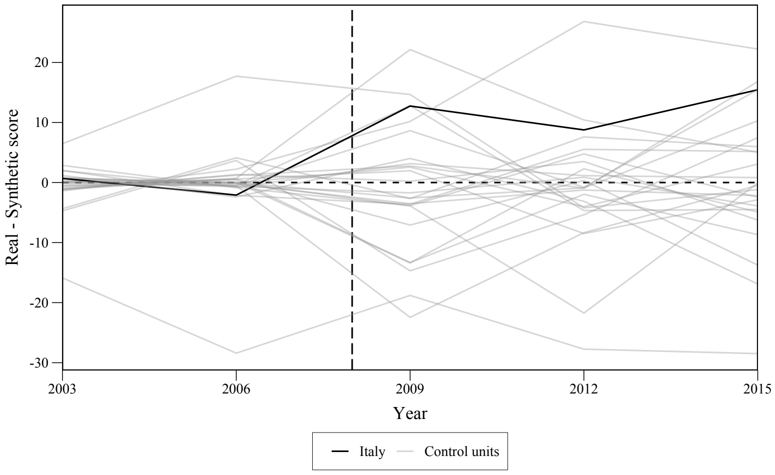
The authors of the paper have not developed a package to run penalized SC yet. However, the `github` page of the paper contains the functions employed in their analysis. I was able to adapt their code to my data, and Figure 11 plots observed/synthetic gaps for treated and control units. The penalized SC assigns positive weights to the USA, Israel, and Russia only (.48, .33, .19, respectively)[51]. Overall, the penalized synthetic control provides additional evidence in favor of the validity of my estimates. Although the p-value is now larger than in baseline estimates, I can still conclude that the Gelmini reform does not seem to negatively impact pupils' achievement.

The countries that are given positive weights in the penalized SC and the one that exploits TIMSS data are different from Italy in many respects. Indeed, the synthetic control is not able to perfectly match Italy's covariate predictors (although it somehow improves on the simple mean of donor pool units). However, the SC is able to match Italy's pre-reform performance with extreme precision, and the high intercorrelations I discussed above suggest that this approach may provide consistent estimates, as long as there are no other large (idiosyncratic) reforms that

[50] This term is given a weight λ (*tuning parameter*), which I set to .1 as in the empirical application carried out by Abadie and L'Hour (2021) (it can also be chosen optimally in a data-driven fashion.).

[51] Unfortunately, small pre-reform RMSEs still make inference based on post- to pre-reform RMSE not a valid solution.

Figure 11: Merged PISA and TIMSS mathematics scores



Note: vertical dashed line denotes the time of the intervention.

impacted either the treated or control units. Moreover, in robustness analysis, I showed that splitting the pre-intervention period into a matching and validation period does not alter qualitative results. Post-reform tracking of reading scores is also reassuring.

The robustness analysis I carried out in this section improves on previous literature which applied SCM to international assessment scores, and strengthens the credibility of the results. In addition to this, it improves on simple regression analysis in that it makes the choice of control units explicit.

8 Conclusion

My thesis attempted to gauge the impact of the 2008-2010 Gelmini school reform on Italian students' educational achievement. The reform aimed at cutting on educational spending by targeting the heaviest budget item: teaching staff. In order to implement these changes, instruction time was reduced in all school cycles, and class size was increased at the margins, among the other interventions. At the same time, the minister Mariastella Gelmini designed a few interventions to boost the efficiency of the education system: for the relevance in this study, the introduction of Invalsi standardized tests in middle school exit exams should be mentioned.

I applied Synthetic Control methods to a panel dataset of six PISA international assessments (waves 2000 to 2015) for 25 countries to carry out a case study of the reform. I found no statistically significant evidence of an impact on reading scores, but large effects on mathematics performance. The inferential strategy based on placebo runs (permutations) set the p-value in a range about the 10% threshold, meaning weak statistical significance.

The staggered implementation of the interventions allowed me to attempt an interpretation of these results: the one intervention that could impact pupils' performance as early as 2009 is the introduction of Invalsi standardized tests in 8th grade. On the contrary, subsequent interventions that cut school resources could only affect later PISA assessments, but there is no evidence of a negative effect on achievement in those waves.

The observed positive effect on math scores may simply be the result of preparation for Invalsi tests and *teaching to the test*: in this regard, PISA tests may not be a reliable measure of pupils' improvements in their (true) knowledge. However, a conservative and economically relevant conclusion can be drawn from my results: the Gelmini school reform did not negatively affect Italian 15 year-olds' achievement in international tests. An explanation for such finding may be the partial implementation of the cuts to school resources, as already suggested by Battistin et al. (2015). Indeed, exploiting PISA questionnaires administered to school principals, I found only modest evidence of negative shocks to student-teacher ratios and shortage of school resources and staff.

My analysis suffers from the scarcity of pre-intervention data, which impacts the choice of weights for control units and covariate predictors. In robustness analysis, I experimented with changing the matching period, the predictors, and the donor pool units. Moreover, I directly addressed the problematic choice of weights by merging TIMSS data (thus increasing pre-reform periods), and applying a recently developed variant of the Synthetic Control that penalizes poor interpolations (see Abadie and L'Hour (2021)). As more PISA waves and further international assessments will become available, the reliability of Synthetic Control methods applied to these data will increase. At the same time, though, more and more reforms and other idiosycratic shocks will hit participating countries, impairing their suitability as control units.

This is one of the few studies that attempted to quantitatively assess the outcome of the Gelmini reform. I found that the reform had only a modest impact on targeted resource items and does not seem to have negatively impacted students' educational achievement; there is also some evidence that it has boosted the efficiency of the system, although this finding could well be a result of *teaching to the test*. Further analysis on the reform is definitely desirable. Moreover, to the best of my knowledge, this is the first attempt to apply Synthetic Controls to a merged PISA-TIMSS database, and the first application of Abadie and L'Hour (2021)'s penalized SCM to international assessments data: by providing remedies to the scarcity of pre-intervention periods, this work opens a new avenue of research in the field of the economics of education.

References

Abadie, Alberto (2021) "Using synthetic controls: Feasibility, data requirements, and methodological aspects," *Journal of Economic Literature*, 59 (2), 391–425.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American statistical Association*, 105 (490), 493–505.

——— (2011) "Synth: An R package for synthetic control methods in comparative case studies," *Journal of Statistical Software*, 42 (13).

——— (2015) "Comparative politics and the synthetic control method," *American Journal of Political Science*, 59 (2), 495–510.

Abadie, Alberto and Javier Gardeazabal (2003) "The economic costs of conflict: A case study of the Basque Country," *American economic review*, 93 (1), 113–132.

Abadie, Alberto and Jérémy L'Hour (2021) "A penalized synthetic control estimator for disaggregated data," *Journal of the American Statistical Association*, 116 (536), 1817–1834.

Acemoglu, Daron, Simon Johnson, Amir Kermani, James Kwak, and Todd Mitton (2016) "The value of connections in turbulent times: Evidence from the United States," *Journal of Financial Economics*, 121 (2), 368–391.

Altinok, Nadir, Claude Diebolt, and Jean-Luc Demeulemeester (2014) "A new international database on education quality: 1965–2010," *Applied Economics*, 46 (11), 1212–1247.

Ammermüller, Andreas (2005) "Educational opportunities and the role of institutions," *ZEW-Centre for European Economic Research Discussion Paper* (05-044).

Anghel, Brindusa, Antonio Cabrales, Jorge Sainz, and Ismael Sanz (2015) "Publicizing the results of standardized external tests: does it have an effect on school outcomes?" *IZA Journal of European Labor Studies*, 4 (1), 1–20.

Angrist, Joshua D and Victor Lavy (1999) "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *The Quarterly journal of economics*, 114 (2), 533–575.

Battistin, Erich, Daniela Vuri, and Simone Schüller (2015) *Bambini che imparano meno? Gli effetti della riforma Gelmini nella scuola primaria*: il Mulino.

Bellei, Cristián (2009) "Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile," *Economics of Education Review*, 28 (5), 629–640.

Belot, Michèle and Vincent Vandenberghe (2011) "Grade retention and educational attainment," *Education Economics*, forthcoming.

Beneito, P and Ó Vicente-Chirivella (2020) "Banning mobile phones at schools: Effects on bullying and academic performance," *Unpublished document.*

Bohn, Sarah, Magnus Lofstrom, and Steven Raphael (2014) "Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population?" *Review of Economics and Statistics*, 96 (2), 258–269.

Boulay, Beth, Beth Gamse, Amy Checkoway, Kenyon Maree, and Lindsay Fox (2011) "Evaluation of Massachusetts Expanded Learning Time (ELT) Initiative: Implementation and Outcomes after Four Years.," *Society for Research on Educational Effectiveness.*

Braga, Michela, Daniele Checchi, and Elena F Meschi (2011) "Institutional reforms and educational attainment in Europe: A long run perspective," *Available at SSRN 1976521*.

Brown, Giorgina, John Micklewright, Sylke V Schnepf, and Robert Waldmann (2005) "Cross-national surveys of learning achievement: How robust are the findings?".

Carneiro, Pedro Manuel and James J Heckman (2003) "Human capital policy."

Caro, Daniel H and Przemysław Biecek (2017) "intsvy: An R package for analyzing international large-scale assessment data," *Journal of Statistical Software*, 81, 1–44.

Cordero, José M, Victor Cristobal, and Daniel Santín (2018) "Causal inference on education policies: A survey of empirical studies using PISA, TIMSS and PIRLS," *Journal of Economic Surveys*, 32 (3), 878–915.

Cottone, Nicoletta [Sole24Ore] (2008) "L'abc della legge Gelmini," `https://st.ilsole24ore.com/art/SoleOnLine4/Norme%20e%20Tributi/2008/10/decreto-scuola-abc-maxiemendamento_2.shtml`.

Figlio, David and Susanna Loeb (2011) "School accountability," *Handbook of the Economics of Education*, 3, 383–421.

Fishbein, Bethany, Pierre Foy, and Liqun Yin (2021) "TIMSS 2019 user guide for the international database."

Fitzpatrick, Maria D (2008) "Starting school at four: The effect of universal pre-kindergarten on children's academic achievement," *The BE Journal of Economic Analysis & Policy*, 8 (1).

Gagliardi, Giovanni [LaRepubblica] (2008) "Studenti, prof e genitori in corteo. Tutti in piazza da Bolzano a Lipari.," `https://www.repubblica.it/2008/10/sezioni/scuola_e_universita/servizi/scuola-2009-4/sciopero-30/sciopero-30.html`.

Garrouste, Christelle (2010) *100 Years of Educational Reforms in Europe: a contextual database.* Ph.D. dissertation, European Commission's Joint Research Centre (JRC).

Goodman, Joshua (2014) "Flaking out: Student absences and snow days as disruptions of instructional time,"Technical report, National Bureau of Economic Research.

Hansen, Benjamin (2011) "School year length and student performance: Quasi-experimental evidence," *Available at SSRN 2269846*.

Hanushek, Eric A and Ludger Wößmann (2006) "Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries," *The Economic Journal*, 116 (510), C63–C76.

——— (2011) "The economics of international differences in educational achievement," *Handbook of the Economics of Education*, 3, 89–200.

——— (2012) "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation," *Journal of economic growth*, 17 (4), 267–321.

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

IlGiornale (2009) "Università, Napolitano: "No tagli indiscriminati". Gelmini: premiamo i migliori e tagliamo sprechi.," `https://www.ilgiornale.it/news/universit-napolitano-no-tagli-indiscriminati-gelmini.html`.

Illiano, Luigi [Sole24Ore] (2008) "Ritorno a scuola, educazione civica in 33 ore.," https://win.gildavenezia.it/docs/Archivio/2008/ago2008/ritorno.htm.

Intravaia, Salvo [LaRepubblica] (2008) "Gelmini: "Così la scuola cambia". Più inglese e "unico maestro".," https://www.orizzontedocenti.it/2008/12/18/gelmini-cosi-la-scuola-cambia-piu-inglese-e-unico-maestro/.

——— (2011) "L'uragano di tagli del govenro sull'istruzione. Il prossimo anno saltano altre 20 mila cattedre.," https://www.repubblica.it/scuola/2011/03/06/news/l_uragano_di_tagli_del_governo_sull_istruzione_il_prossimo_anno_saltano_altre_20_mila_cattedre-13247197/.

Jahanshahi, Babak and Arash Naghavi (2017) "Education reform and education gaps," *Applied Economics Letters*, 24 (19), 1385–1388.

James-Burdumy, Susanne, Mark Dynarski, Mary Moore, John Deke, Wendy Mansfield, Carol Pistorino, and Elizabeth Warner (2005) "When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program. Final Report.," *US Department of Education*.

Jennings, Jennifer L and Jonathan Marc Bearak (2014) ""Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance," *Educational Researcher*, 43 (8), 381–389.

Johnson, Clay Stephen (2013) "Compared to What? The Effectiveness of Synthetic Control Methods for Causal Inference in Educational Assessment."

LaRepubblica (2008) "Le università in rivolta. Domani il corteo dei Cobas.," https://www.repubblica.it/2008/09/sezioni/scuola_e_universita/servizi/universita-2009/protesta-16-ott/protesta-16-ott.html.

——— (2009) "Napolitano: "No tagli a ricerca". Gelmini: "Elimiano solo gli sprechi".," https://www.repubblica.it/2009/01/sezioni/scuola_e_universita/servizi/universita-2009-1/naolitano-critica/naolitano-critica.html.

Lavy, Victor (2012) "Expanding school resources and increasing time on task: Effects of a policy experiment in Israel on student academic achievement and behavior,"Technical report, National Bureau of Economic Research.

——— (2015) "Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries," *The Economic Journal*, 125 (588), F397–F424.

Lavy, Victor and Analía Schlosser (2005) "Targeted Remedial Education for Underperforming Teenagers: Costs and Benefits," *Journal of Labor Economics*, 23 (4), 839–874.

Lazear, Edward P (2006) "Speeding, terrorism, and teaching to the test," *The Quarterly Journal of Economics*, 121 (3), 1029–1061.

Lee, Jong-Wha and Robert J Barro (2001) "Schooling quality in a cross–section of countries," *Economica*, 68 (272), 465–488.

Longo, Christopher (2010) "Fostering creativity or teaching to the test? Implications of state testing on the delivery of science instruction," *The Clearing House*, 83 (2), 54–57.

Mandel, Philipp and Bernd Süssmuth (2011) "Total instructional time exposure and student achievement: An extreme bounds analysis based on German state-level variation."

Marcotte, Dave E and Steven W Hemelt (2008) "Unscheduled school closings and student performance," *Education Finance and Policy*, 3 (3), 316–338.

Meyer, Erik and Chris Van Klaveren (2013) "The effectiveness of extended day programs: Evidence from a randomized field experiment in the Netherlands," *Economics of Education Review*, 36, 1–11.

Murnane, Richard, John B Willett, and Frank Levy (1995) "The growing importance of cognitive skills in wage determination."

Nickell, Stephen (2004) "Poverty and worklessness in Britain," *The Economic Journal*, 114 (494), C1–C25.

Peri, Giovanni and Vasil Yasenov (2015) "The labor market effects of a refugee wave: Applying the synthetic control method to the Mariel boatlift,"Technical report, National Bureau of Economic Research.

Pinotti, Paolo (2015) "The economic costs of organised crime: Evidence from Southern Italy," *The Economic Journal*, 125 (586), F203–F232.

Pischke, Jörn-Steffen (2007) "The impact of length of the school year on student performance and earnings: Evidence from the German short school years," *The Economic Journal*, 117 (523), 1216–1242.

Rivkin, Steven G and Jeffrey C Schiman (2015) "Instruction time, classroom quality, and academic achievement," *The Economic Journal*, 125 (588), F425–F448.

Rockoff, Jonah and Lesley J Turner (2010) "Short-run impacts of accountability on school quality," *American Economic Journal: Economic Policy*, 2 (4), 119–47.

Rosenbaum, Paul R (2005) "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *The American Statistician*, 59 (2), 147–152.

Rubin, Donald B (1974) "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of educational Psychology*, 66 (5), 688.

Schleicher, Andreas (2019) "PISA 2018: Insights and Interpretations.," *oecd Publishing.*

Schneeweis, Nicole (2011) "Educational institutions and the integration of migrants," *Journal of Population Economics*, 24 (4), 1281–1308.

Schütz, Gabriela, Heinrich W Ursprung, and Ludger Wößmann (2008) "Education policy and equality of opportunity," *Kyklos*, 61 (2), 279–308.

della Sera, Corriere (2008) "Scuola: l'ora di lezione da 50 a 60 minuti, licei e tecnici: da 714 scendono a venti," `https://www.corriere.it/cronache/08_dicembre_18/scuola_decreti_attuativi_89435804-ccf8-11dd-95df-00144f02aabc.shtml`.

Sims, David P (2008) "Strategic responses to school accountability measures: It's all in the timing," *Economics of Education Review*, 27 (1), 58–68.

Soh, Yew Chong, Ximena V Del Carpio, and Liang Choon Wang (2021) "The Impact of Language of Instruction in Schools on Student Achievement."

Sole24Ore (2010) "Studenti in piazza e università occupate per protestare contro la riforma Gelmini.," `https://st.ilsole24ore.com/art/notizie/2010-10-15/studenti-piazza-universita-occupate-093605.shtml`.

Vitali, Alessandra [LaRepubblica] (2008) "Roma invasa: "Siamo un milione". E i ragazzi circondano il ministero.," `https://www.repubblica.it/2008/10/sezioni/scuola_e_universita/servizi/scuola-2009-4/manifestazione-roma/manifestazione-roma.html`.

Volante, Louis and Don A Klinger (2021) "PISA and Education Reform in Europe: Cases of policy inertia, avoidance, and refraction," *European Education*, 53 (1), 45–56.

Wößmann, Ludger et al. (2015) *Universal basic skills what countries stand to gain:*

What countries stand to gain: OECD publishing.

Wößmann, Ludger (2003) "Schooling resources, educational institutions and student performance: the international evidence," *Oxford bulletin of economics and statistics*, 65 (2), 117–170.

——— (2016) "The importance of school systems: Evidence from international differences in student achievement," *Journal of Economic Perspectives*, 30 (3), 3–32.

Zimmer, Ron, Laura Hamilton, and Rachel Christina (2010) "After-school tutoring in the context of no child left behind: Effectiveness of two programs in the Pittsburgh public schools," *Economics of education Review*, 29 (1), 18–28.

Appendix A. Countries excluded from the donor pool

As anticipated in section 2, the original donor pool included Australia, Austria, Belgium, Bulgaria, Canada, Croatia, Czech Republic, Denmark, England, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Luxembourg, the Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Spain, Sweden, Switzerland, Turkey, the USA. The final donor pool includes Australia, Austria, Belgium, Bulgaria, Canada, Czech Republic, Denmark, England, Finland, France, Germany, Iceland, Ireland, Israel, Italy, Latvia, Luxembourg, theNetherlands, New Zealand, Norway, Russia, Spain, Sweden, Switzerland, the USA.

Here, I show which countries were excluded and why. The general rationale behind these exclusions is that of avoiding including countries that experienced large idiosyncratic shocks that may bias synthetic control estimates. I did not exclude countries where the education policy was fluid during the period under consideration. I carefully reviewed the Eurydice database to explore potential reforms that may cause concerning shocks; additional online sources were consulted for those countries that are not covered by Eurybase.

Croatia's single structure education is regulated by the Act on Education in Primary and Secondary Schools (first initiated in 2008). The National Pedagogical Standards for Single Structure Education System, adopted in 2008, set a firm base for further development and work in primary schools and criteria regarding, for example, class size and number of classes in schools. Further regislation introduced new curricula since 2009.

Greece introduced optional primary and pre-primary all-day school with Law 2525/97. Moreover, changes in primary and pre-primary curricula, class size and schedule occured in 2003, 2006 and 2007; similarly, new curricula for lower secondary school were introduced in 2003 and new lower-secondary timetables in 2012 and 2013; new class sizes in lower-secondary school in 2013.

Hungary carried out its first National Assessment of Basic Competences (NABC) in 2001 (implemented since 2007) with the primary aim of assessing the effectiveness of schools; the NABC is an annual assessment system which covers almost all students in grades 6, 8 and 10 and tests pupils in reading and mathematic literacy.

Japan implemented a number of reforms for compulsory education at the beginning of the new millennium, including the reduction of school hours (primarily through eliminating Saturday classes). Shortly after the implementation of the new curriculum, Japanese students declined in their achievement on international comparative tests. In response, a number of gradual changes were implemented, designed to help improve the academic performance of Japanese students while maintaining some of the benefits of the earlier reforms. Key changes included a gradual increase in the required topics to be taught in the standard academic subjects, a gradual increase in the number of hours devoted to these subjects, and the implementation of national standardized testing at the end of the 6th and 9th grades (National Survey on Educational Performance) in mathematics and Japanese, starting in 2007. Further reforms of the curricula and school hours in 2011.

Korea has shown what can be done to improve education over the last two decades. It has reduced class sizes (2004 "7.20" Initiative) and extended schooling hours to meet a surging demand for better education. To ensure quality control of the national curriculum, the National Assessment of Education Achievement (NAEA) is conducted annually since 2008. The goal of the NAEA is to assess educational progress and achievement nationwide. The NAEA is administered to all students in Grades 9 and 11, in the subjects of Korean, mathematics, and English. Students' results on the NAEA do not impact their school grades; schools do, however, provide appropriate support for students based on their NAEA results. Primary and secondary schools (including middle schools) do not hold Saturday classes since 2012.

Poland implemented a school education reform in September 1999 that extended the duration of full-time compulsory education by one year as part of a new structure establishing a 6-year primary school and a 3-year lower secondary school. Since the s.y. 1999/2000, all pupils who finished the 6-year primary school continued their education in the 3-year lower secondary school; at the end of lower secondary education, pupils took a compulsory external exam. The school education act also reformed upper secondary tracks. In September 2004, one-year compulsory preschool preparatory classes were introduced for 6-year-old children.

Portugal has put in place a vast set of policies designed to improve learning outcomes since 2005, especially for disadvantaged people: the Government has devoted more resources to supporting children from low income families and a new system of evaluating teachers and schools has been put in place to increase accountability. As part of the reform, all students in 4th, 6th and 9th grades take part in annual national assessments in Portuguese language and mathematics. In particular, they introduced standardized national exams for 9th graders since s.y. 2004/5 (moreover, low-stakes tests for 6th graders are not anonymous since Legal

Norm 2351/07, and are now administered to everyone, not samples). Unsurprisingly, PISA reports show a boost in achievement since 2009, most likely through reductions in repetition rates.

Romania's general legal framework to organise, administrate and provide education is established through the Constitution and the Law of National Education (Law 1/2011): it establishes, for instance, class sizes in lower secondary education. Since 2010, at the end of grade 6, all schools organise and conduct a student assessment called "The National Assessment at the End of Grade 6", with two cross-disciplinary tests. The students who complete grade 8 participate in a similar national examination which is used as a summative assessment of competences acquired throughout lower secondary education (the average general mark obtained at the National Evaluation is one of the criteria for admission to public high-school education).

Serbia implemented the education system of the Socialist Federative Republic of Yugoslavia until 2005. The 2009 Law on Foundations of the Education System puts emphasis on the prohibition of discrimination and segregation, and introduces a new assessment and evaluation policy. By s.y. 2007/08, all study programmes at all higher education institutions had been reformed. Summative assessment is used at the end of the 8th grade, when pupils are required to take the final exam in order to complete basic education. The realisation of standards of pupils' achievement is also examined through the national examination on a sample of schools and pupils in different grades. This examination affects neither pupils' grades nor secondary school enrolment; however, pupils are required to participate.

Slovakia has started a reform of the public administration and the organization of schools (management, establishment and dissolution of schools and school facilities, and self-governance of schools) with Act 596/03 of the Law Code. In 2008, the new Education Act, which has created legislative space for the implementation of the reform of regional education, was approved: the major changes include reduction of the number of pupils in classes, changes in the curricula, and the implementation of a new philosophy of education which is more child-oriented compared to performance-oriented or contents-oriented education of the past. The details about the organization of the school year are laid down in the Education Act, but exact dates are annually publicized in School-Year's Guide. Organisation and provision of the educational process at school is regulated, apart by other, by the Decree on primary schools (Decree 231/09).

Turkey's Basic education Programme (BeP) started in 1998 and had an impact on almost all students, expanding primary school education, improving the quality of education and overall student outcomes, etc. One of the major changes introduced with the BeP programme involved the compulsory education law: this change was first implemented in the 1997/98 school year, and in 2003 the first students graduated from the eight-year compulsory education system. Since the launch of this programme, the attendance rate among students within the eight-year primary education system increased from around 85% to nearly 100%, while the attendance rate in pre-primary programmes increased from 10% to 25%. In addition, the system was expanded to include 3.5 million more pupils, and average class size was reduced to roughly 30. In line with these changes, new curricula were implemented in the 2006-2007 school year and PISA 2009 students had already been taught for one year using the new curriculum. Several projects implemented in Turkey over the past decade have addressed equity issues. Private investments were also used to increase the capacity of the school system in the country.

Appendix B - R Code

```
# PACKAGES ----

#install_github("pbiecek/PISA2000lite")
#install_github("pbiecek/PISA2003lite")
#install_github("pbiecek/PISA2006lite")
#devtools::install_github("pbiecek/PISA2009lite")
#install_github("pbiecek/PISA2012lite")
#devtools::install_github("edunford/tidysynth")
library(Synth)
require(tidysynth)
library(ddpcr)
library(tidyverse)
library(plyr)
library(dplyr)
library(foreign)
library(ggplot2)
library(Hmisc)
library(memisc)
library(reshape)
library(intsvy)
library(devtools)
library(haven)
library(stringr)
library(labelled)
library(countrycode)
library(readxl)
library(tibble)
library(janitor)
library(imputeTS)
library(VIM)
library(data.table)
library(PISA2000lite)
library(PISA2003lite)
library(PISA2006lite)
library(PISA2009lite)
library(PISA2012lite)
```

```
# FUNCTIONS ----
setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/Functions")
source("GraphFunctions.R")


# IMPORT DATASETS ----
setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/P00")
pisa00stumath <- PISA2000lite::math2000
pisa00read <- PISA2000lite::read2000
pisa00sch <- PISA2000lite::school2000
pisa00math <- merge(pisa00stumath,pisa00sch,by=c("CNT", "
   SCHOOLID"))
rm(pisa00stumath, pisa00sch)
#write.table(colnames(pisa00math),file="var.txt")
#Useful variables: (student) "ST37Q01", ("FISCED"+"MISCED"),
   "ST16Q01", (school)
# "SC11Q04", "SC01Q01", "STRATIO", "SC22Q01", "SC22Q05", "
   SC03Q01", "SC04Q01",
# "SC18Q04", SC17Q03, "SCHAUTON", No external monitoring

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/P03")
pisa03stu <- PISA2003lite::student2003
pisa03sch <- PISA2003lite::school2003
pisa03 <- merge(pisa03stu,pisa03sch,by=c("CNT", "SCHOOLID"))
rm(pisa03stu, pisa03sch)
#write.table(colnames(pisa03),file="var.txt")
#Useful variables: (student) "ST19Q01", "HISCED", ["IMMIG"] "
   ST15Q01", (school) "SC08Q09", "SC01Q01",
# "STRATIO", "SC26Q01", "SC26Q05", "SC03Q01", "SC04Q01", "
   SC13Q04", no tracking by auth. ("SC27Q07"??),
# "SCHAUTON", "SC20Q04"

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/P06")
pisa06stu <- PISA2006lite::student2006
pisa06sch <- PISA2006lite::school2006
pisa06 <- merge(pisa06stu,pisa06sch,by=c("COUNTRY", "SCHOOLID
   "))
```

```
rm(pisa06stu, pisa06sch)
#write.table(colnames(pisa06),file="var.txt")
#Useful variables: (student) "ST15Q01", "HISCED", ["IMMIG"] "
   ST11Q01", (school) "SC14Q08", "SC07Q01",
# "STRATIO", "SC11QA1", "SC11QA2", "SC11QA3", "SC11QA4", "
   SC11QE1", "SC11QE2", "SC11QE3",
# "SC11QE4", "SC02Q01", "SC03Q01", "SC15Q02", "SC17Q05", ???
   SCHAUTON, No external monitoring

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/P09")
pisa09stu <- PISA2009lite::student2009
pisa09sch <- PISA2009lite::school2009
pisa09 <- merge(pisa09stu,pisa09sch,by=c("CNT", "SCHOOLID"))
#write.table(colnames(pisa09),file="var.txt")
#Useful variables: (student) "ST22Q01", "HISCED", ["IMMIG"] "
   ST17Q01", (school) "SC11Q08",
# "SC04Q01", "STRATIO", "SC24QA1", "SC24QA2", "SC24QA3", "
   SC24QA4", "SC24QA5",
# "SC24QE1", "SC24QE2", "SC24QE3", "SC24QE4", "SC24QE5", "
   SC02Q01", "SC03Q01",
# "SC16Q04", "SC22Q05", ???SCHAUTON, "SC23Q04"

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/P12")
pisa12stu <- PISA2012lite::student2012
pisa12sch <- PISA2012lite::school2012
pisa12 <- merge(pisa12stu,pisa12sch,by=c("CNT", "SCHOOLID"))
#quiet(data("student2012dict", package = "PISA2012lite"))
#quiet(data("school2012dict", package = "PISA2012lite"))
#quiet(data("parent2012dict", package = "PISA2012lite"))
#write.table(student2012dict, file="p12dictstu.txt")
#write.table(school2012dict, file="p12dictsch.txt")
#write.table(parent2012dict, file="p12dictpar.txt")
#rm(student2012dict,school2012dict,parent2012dict)
#grep(student2012dict, pattern = "keyword", value = TRUE)
#table(student2012[["ST28Q01"]])
#Useful variables: (student) "ST28Q01", "HISCED", ["IMMIG"] "
   ST20Q01", (school) "SC14Q06", "SC03Q01",
# STRATIO", "SC33Q01A", "SC33Q01B", "SC33Q01C", "SC33Q01D", "
   SC33Q01E", "SC33Q05A",
# "SC33Q05B", "SC33Q05C", "SC33Q05D", "SC33Q05E", "SC01Q01",
   "SC02Q01", "SC18Q04",
# "SC19Q02", "SCHAUTON", "SC30Q04"

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/P15")
#pisa.var.label(folder=getwd(), school.file="CY07_MSU_SCH_QQQ
   .sav", student.file="CY6_MS_CMB_STU_QQQ.sav")
pisa15 <- pisa.select.merge(folder=getwd(),
school.file="CY6_MS_CMB_SCH_QQQ.sav",
student.file="CY6_MS_CMB_STU_QQQ.sav",
student=c("ST013Q01TA", "HISCED", "IMMIG", "ST019AQ01T"),
school=c("SC017Q05NA", "SC017Q01NA", "SC001Q01TA", "STRATIO",
"SC010Q01TA", "SC010Q01TB", "SC010Q01TC",
"SC010Q01TD", "SC010Q01TE", "SC010Q05TA",
"SC010Q05TB", "SC010Q05TC", "SC010Q05TD",
"SC010Q05TE", "SC013Q01TA", "SC016Q01TA",
"SC035Q05TA", "SC036Q02TA", "SCHAUT", "SC032Q04TA"),
countries=c("AUS", "AUT", "BEL", "BGR", "CAN","CHE", "CZE", "
   DEU", "DNK",
"ESP", "EST" ,"FIN", "FRA" ,"GBR" ,"GRC", "HUN", "HRV",
"IRL" ,"ISL" ,"ISR", "ITA" ,"JPN" , "KOR", "LTU" ,"LUX",
"LVA" ,"NLD", "NOR" ,"NZL" , "POL" , "PRT", "ROU","RUS" ,
"SRB","SVK" , "SVN", "SWE","TUR", "UKR", "USA"))
#quiet(pisa.var.label(folder=getwd(), school.file="CY6_MS_CMB
   _SCH_QQQ.sav", student.file="CY6_MS_CMB_STU_QQQ.sav"))

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/P18")
#pisa.var.label(folder=getwd(), school.file="CY07_MSU_SCH_QQQ
   .sav", student.file="CY07_MSU_STU_QQQ.sav")
pisa18 <- pisa.select.merge(folder=getwd(),
school.file="CY07_MSU_SCH_QQQ.sav",
student.file="CY07_MSU_STU_QQQ.sav",
student=c("ST013Q01TA", "HISCED", "IMMIG", "ST019AQ01T"),
school=c("SC017Q05NA", "SC017Q01NA", "SC001Q01TA", "STRATIO",
    "SC013Q01TA",
"SC016Q01TA", "SC154Q05WA", "SC036Q02TA"),
countries=c("AUS", "AUT", "BEL","BGR","CAN","CHE", "CZE", "
   DEU",
"DNK", "ESP", "EST", "FIN", "FRA", "GBR",
"GRC", "HUN", "HRV","IRL", "ISL", "ISR", "ITA",
"JPN", "KOR", "LTU", "LUX", "LVA", "NLD",
"NOR", "NZL", "POL", "PRT","ROU","RUS","SRB", "SVK", "SVN",
"SWE","TUR", "UKR", "USA"))
#Also look at questionnaire question SC037Q02TA.
#quiet(pisa.var.label(folder=getwd(), school.file="CY07_MSU_
   SCH_QQQ.sav", student.file="CY07_MSU_STU_QQQ.sav"))


setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/T99")
#timssg8.var.label(folder= getwd())
timss99 <- timssg8.select.merge(folder= getwd(),
countries=c("AUS", "BFL", "BGR","CAN","CZE", "ENG", "EST","
   FIN",
"HUN", "ISR","ITA", "JPN", "KOR", "LTU", "LVA", "NLD", "NOR",
"NZL", "ROU","RUS", "SVK", "SVN", "SWE", "USA"),
student =c("BSBGBOOK", "BSDGEDUP", "BSBGBRN1"),
school=c("BCBGST01", "BCBGCOMM", "BCDGSTRA", "BCBGRP01", "
   BCBGRP04", "BCBSST18"))
#quiet(timssg8.var.label(folder= getwd()))

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/T03")
#timssg8.var.label(folder= getwd())
timss03 <- timssg8.select.merge(folder= getwd(),
countries=c("AUS", "BFL","BGR", "ENG", "EST", "HUN", "ISR",
"ITA", "JPN", "KOR", "LTU", "LVA", "NLD", "NOR",
"NZL", "ROU","RUS", "SVK", "SVN", "SWE", "USA"),
student =c("BSBGBOOK", "BSDGEDUP", "BSBGBORN"),
school=c("BCBGST01", "BCBGCOMU", "BCBGSH18"))
#quiet(timssg8.var.label(folder= getwd()))

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/T07")
#timssg8.var.label(folder= getwd())
timss07 <- timssg8.select.merge(folder= getwd(),
countries=c("AUS","BGR", "CZE", "ENG", "HUN", "ISR",
"ITA", "JPN", "KOR", "LTU", "NOR","ROU",
"RUS", "SVN", "SWE", "UKR", "USA"),
student =c("BS4GBOOK", "BSDGEDUP", "BS4GBORN"),
school=c("BC4GST01", "BC4GCOMU", "BC4GSH18"))
#quiet(timssg8.var.label(folder= getwd()))

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/T11")
#timssg8.var.label(folder= getwd())
timss11 <- timssg8.select.merge(folder= getwd(),
countries=c("AUS", "ENG", "FIN", "HUN", "ISR",
"ITA", "JPN", "KOR", "LTU", "NOR", "NZL",
"ROU","RUS", "SVN", "SWE", "UKR", "USA"),
student =c("BSBG04", "BSDGEDUP", "BSBG09A"),
school=c("BCBG09AA", "BCBG05B", "BCBG09BA"))
#quiet(timssg8.var.label(folder= getwd()))

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/T15")
#timssg8.var.label(folder= getwd())
timss15 <- timssg8.select.merge(folder= getwd(),
countries=c("AUS", "CAN", "ENG", "HUN", "IRL", "ISR",
"ITA", "JPN", "KOR", "LTU", "NOR", "NZL",
"RUS", "SVN", "SWE", "USA"),
student =c("BSBG04", "BSDGEDUP", "BSBG10A"),
school=c("BCBG13AA", "BCBG05B", "BCBG13BA"))
#quiet(timssg8.var.label(folder= getwd()))

setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/T19")
#timssg8.var.label(folder= getwd())
timss19 <- timssg8.select.merge(folder= getwd(),
countries=c("AUS", "ENG", "FIN", "FRA", "HUN", "IRL", "ISR",
"ITA", "JPN", "KOR", "LTU", "NOR", "NZL", "PRT",
"ROU","RUS", "SWE", "USA"),
student = c("BSBG04", "BSDGEDUP", "BSBG09A"),
school = c("BCBG13AA", "BCBG05B", "BCBG13BA"))
#quiet(timssg8.var.label(folder= getwd()))

#GDPpc data from World Bank (GDP per capita, PPP, constant
   2017 international $)
gdp <- read.csv("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/
   Master's␣Thesis/Data␣analysis/Predictors'␣data/GDP.csv",
   header=TRUE)
```

```
names(gdp)[names(gdp) == 'Country.Code'] <- 'CNT'
gdp <- gdp[-c(26:30),-c(1,3,4)]
names(gdp)[2:27] <- 1995:2020
gdp <- gdp[,which(names(gdp) %in% c("CNT"
   ,1999,2000,2003,2006,2007,2009,2011,2012,2015,2018,2019))]
gdp <- reshape(gdp,idvar="CNT",varying=list(2:12), v.names="
   GDPpc",
times=c
   (1999,2000,2003,2006,2007,2009,2011,2012,2015,2018,2019),
   direction="long")
gdp <- gdp[order(gdp$CNT,gdp$time),]
gdp$GDPpc <- as.numeric(gdp$GDPpc)

#Private schools data from World Bank (School enrollment,
   primary, private, as % of total primary)
private <- read.csv("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/
   Master's␣Thesis/Data␣analysis/Predictors'␣data/private.csv
   ", header=TRUE)
names(private)[names(private) == 'Country.Code'] <- 'CNT'
private <- private[-c(26:30),-c(1,3,4)]
names(private)[2:27] <- 1995:2020
private <- reshape(private,idvar="CNT",varying=list(2:27), v.
   names="Private",
times=1995:2020,direction="long")
private <- private[order(private$CNT,private$time),]
private$Private <- as.numeric(private$Private)
#Imputation of missing data (25 is the # of countries, 26 the
    # of years)
for (i in 1:25) {
        private$Private[((1+26*(i-1)):(26+26*(i-1)))] <- na_
           interpolation(private$Private[((1+26*(i-1)):(26+26
           *(i-1)))],
        option = "linear")
}
private <- private[private$time %in% c
   (1999,2000,2003,2006,2007,2009,2011,2012,2015,2018,2019),]
rm(i)


# COMPUTE COUNTRY MEANS/SD FOR SCORES AND PREDICTORS ----
```

```
#see demo(PISA2006lite)
#Pisa 2000
pisa00math$CNT <- tolower(pisa00math$CNT)
pisa00math$CNT <- gsub("(^|[[:space:]])([[:alpha:]])", "\\1\\
   U\\2", pisa00math$CNT, perl = TRUE)
for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa00math[,c(paste("PV",i,"MATH",sep="")
           ,"W_FSTUWT")],
        pisa00math[,"CNT"], function(x) weighted.mean(x[,1],
           x[,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa00math[,c(paste("PV",i,"MATH",
           sep=""),"W_FSTUWT")],
        pisa00math[,"CNT"], function(x) wtd.var(x[,1], x[,2])
           ))))))}
MeansP00 <- (means1+means2+means3+means4+means5)/5
SdMP00 <- (sd1+sd2+sd3+sd4+sd5)/5
MeansP00 <- as.data.frame(MeansP00)
SdMP00 <- as.data.frame(SdMP00)
MeansP00 <- tibble::rownames_to_column(MeansP00, "CNT")
MeansP00$CNT <- gsub("␣", "", MeansP00$CNT, fixed = TRUE)
SdMP00 <- tibble::rownames_to_column(SdMP00, "CNT")
SdMP00$CNT <- gsub("␣", "", SdMP00$CNT, fixed = TRUE)

pisa00read$CNT <- tolower(pisa00read$CNT)
pisa00read$CNT <- gsub("(^|[[:space:]])([[:alpha:]])", "\\1\\
   U\\2", pisa00read$CNT, perl = TRUE)
for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa00read[,c(paste("PV",i,"READ",sep="")
           ,"W_FSTUWT")],
        pisa00read[,"CNT"], function(x) weighted.mean(x[,1],
           x[,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa00read[,c(paste("PV",i,"READ",
           sep=""),"W_FSTUWT")],
        pisa00read[,"CNT"], function(x) wtd.var(x[,1], x[,2])
           ))))))}
Read00 <- (means1+means2+means3+means4+means5)/5
SdRP00 <- (sd1+sd2+sd3+sd4+sd5)/5
Read00 <- as.data.frame(Read00)
SdRP00 <- as.data.frame(SdRP00)
```

```
remove(means1, means2, means3, means4, means5, sd1, sd2, sd3,
    sd4, sd5)
Read00 <- tibble::rownames_to_column(Read00, "CNT")
Read00$CNT <- gsub("␣", "", Read00$CNT, fixed = TRUE)
SdRP00 <- tibble::rownames_to_column(SdRP00, "CNT")
SdRP00$CNT <- gsub("␣", "", SdRP00$CNT, fixed = TRUE)
#Austria's true reading value was 492, instead of 507 (as in
   the first PISA report and in this data)
Read00[Read00$CNT=="Austria",2] <- 492

pisa00math$ST37Q01 <- dplyr::recode(pisa00math$ST37Q01, "None
   "=0L, "1-10"=0L, "11-50"=0L, "51-100"=0L,
"101-250"=1L, "251-500"=1L, "More␣than␣500"=1L)
BookP00 <- unclass(by(pisa00math[,c("ST37Q01","W_FSTUWT")],
   pisa00math[,"CNT"], function(x) weighted.mean(x[,1], x
   [,2], na.rm=TRUE)))
SdBookP00 <- sqrt(unclass(by(pisa00math[,c("ST37Q01","W_
   FSTUWT")], pisa00math[,"CNT"], function(x) wtd.var(x[,1],
   x[,2], na.rm=TRUE))))
BookP00 <- as.data.frame(BookP00)
BookP00 <- tibble::rownames_to_column(BookP00, "CNT")
BookP00$CNT <- gsub("␣", "", BookP00$CNT, fixed = TRUE)
SdBookP00 <- as.data.frame(SdBookP00)
SdBookP00 <- tibble::rownames_to_column(SdBookP00, "CNT")
SdBookP00$CNT <- gsub("␣", "", SdBookP00$CNT, fixed = TRUE)

pisa00math$ST16Q01 <- dplyr::recode(pisa00math$ST16Q01, "<
   Country␣of␣Test>"=0L, "Other"=1L)
ImmigP00 <- unclass(by(pisa00math[,c("ST16Q01","W_FSTUWT")],
   pisa00math[,"CNT"], function(x) weighted.mean(x[,1], x
   [,2], na.rm=TRUE)))
SdImmigP00 <- sqrt(unclass(by(pisa00math[,c("ST16Q01","W_
   FSTUWT")], pisa00math[,"CNT"], function(x) wtd.var(x[,1],
   x[,2], na.rm=TRUE))))
ImmigP00 <- as.data.frame(ImmigP00)
ImmigP00 <- tibble::rownames_to_column(ImmigP00, "CNT")
ImmigP00$CNT <- gsub("␣", "", ImmigP00$CNT, fixed = TRUE)
SdImmigP00 <- as.data.frame(SdImmigP00)
SdImmigP00 <- tibble::rownames_to_column(SdImmigP00, "CNT")
SdImmigP00$CNT <- gsub("␣", "", SdImmigP00$CNT, fixed = TRUE)
```

```
pisa00math$SC11Q04 <- dplyr::recode(pisa00math$SC11Q04, "Not␣
   at␣all"=0L, "A␣little"=0L, "Some"=1L, "A␣lot"=1L)
ShortP00 <- unclass(by(pisa00math[,c("SC11Q04","W_FSTUWT")],
   pisa00math[,"CNT"], function(x) weighted.mean(x[,1], x
   [,2], na.rm=TRUE)))
SdShortP00 <- sqrt(unclass(by(pisa00math[,c("SC11Q04","W_
   FSTUWT")], pisa00math[,"CNT"], function(x) wtd.var(x[,1],
   x[,2], na.rm=TRUE))))
ShortP00 <- as.data.frame(ShortP00)
ShortP00 <- tibble::rownames_to_column(ShortP00, "CNT")
ShortP00$CNT <- gsub("␣", "", ShortP00$CNT, fixed = TRUE)
SdShortP00 <- as.data.frame(SdShortP00)
SdShortP00 <- tibble::rownames_to_column(SdShortP00, "CNT")
SdShortP00$CNT <- gsub("␣", "", SdShortP00$CNT, fixed = TRUE)

pisa00math$SC21Q03 <- dplyr::recode(pisa00math$SC21Q03, "Not␣
   at␣all"=0L, "A␣little"=0L, "Some"=1L, "A␣lot"=1L)
TeachP00 <- unclass(by(pisa00math[,c("SC21Q03","W_FSTUWT")],
   pisa00math[,"CNT"], function(x) weighted.mean(x[,1], x
   [,2], na.rm=TRUE)))
SdTeachP00 <- sqrt(unclass(by(pisa00math[,c("SC21Q03","W_
   FSTUWT")], pisa00math[,"CNT"], function(x) wtd.var(x[,1],
   x[,2], na.rm=TRUE))))
TeachP00 <- as.data.frame(TeachP00)
TeachP00 <- tibble::rownames_to_column(TeachP00, "CNT")
TeachP00$CNT <- gsub("␣", "", TeachP00$CNT, fixed = TRUE)
SdTeachP00 <- as.data.frame(SdTeachP00)
SdTeachP00 <- tibble::rownames_to_column(SdTeachP00, "CNT")
SdTeachP00$CNT <- gsub("␣", "", SdTeachP00$CNT, fixed = TRUE)

pisa00math$SC03Q01 <- dplyr::recode(pisa00math$SC03Q01, "
   Public"=0L, "Private"=1L)
PrivateP00 <- unclass(by(pisa00math[,c("SC03Q01","W_FSTUWT")
   ], pisa00math[,"CNT"], function(x) weighted.mean(x[,1], x
   [,2], na.rm=TRUE)))
SdPrivateP00 <- sqrt(unclass(by(pisa00math[,c("SC03Q01","W_
   FSTUWT")], pisa00math[,"CNT"], function(x) wtd.var(x[,1],
   x[,2], na.rm=TRUE))))
PrivateP00 <- as.data.frame(PrivateP00)
PrivateP00 <- tibble::rownames_to_column(PrivateP00, "CNT")
PrivateP00$CNT <- gsub("␣", "", PrivateP00$CNT, fixed = TRUE)
```

```
SdPrivateP00 <- as.data.frame(SdPrivateP00)
SdPrivateP00 <- tibble::rownames_to_column(SdPrivateP00, "CNT
   ")
SdPrivateP00$CNT <- gsub("␣", "", SdPrivateP00$CNT, fixed =
   TRUE)

pisa00math$SC18Q04 <- dplyr::recode(pisa00math$SC18Q04, "Yes"
   =1L, "No"=0L)
AccountP00 <- unclass(by(pisa00math[,c("SC18Q04","W_FSTUWT")
   ], pisa00math[,"CNT"], function(x) weighted.mean(x[,1], x
   [,2], na.rm=TRUE)))
SdAccountP00 <- sqrt(unclass(by(pisa00math[,c("SC18Q04","W_
   FSTUWT")], pisa00math[,"CNT"], function(x) wtd.var(x[,1],
   x[,2], na.rm=TRUE))))
AccountP00 <- as.data.frame(AccountP00)
AccountP00 <- tibble::rownames_to_column(AccountP00, "CNT")
AccountP00$CNT <- gsub("␣", "", AccountP00$CNT, fixed = TRUE)
SdAccountP00 <- as.data.frame(SdAccountP00)
SdAccountP00 <- tibble::rownames_to_column(SdAccountP00, "CNT
   ")
SdAccountP00$CNT <- gsub("␣", "", SdAccountP00$CNT, fixed =
   TRUE)

StratioP00 <- unclass(by(pisa00math[,c("STRATIO","W_FSTUWT")
   ], pisa00math[,"CNT"], function(x) weighted.mean(x[,1], x
   [,2], na.rm=TRUE)))
SdStratioP00 <- sqrt(unclass(by(pisa00math[,c("STRATIO","W_
   FSTUWT")], pisa00math[,"CNT"], function(x) wtd.var(x[,1],
   x[,2], na.rm=TRUE))))
StratioP00 <- as.data.frame(StratioP00)
StratioP00 <- tibble::rownames_to_column(StratioP00, "CNT")
StratioP00$CNT <- gsub("␣", "", StratioP00$CNT, fixed = TRUE)
SdStratioP00 <- as.data.frame(SdStratioP00)
SdStratioP00 <- tibble::rownames_to_column(SdStratioP00, "CNT
   ")
SdStratioP00$CNT <- gsub("␣", "", SdStratioP00$CNT, fixed =
   TRUE)

P00 <- list(MeansP00, BookP00, ImmigP00, ShortP00, TeachP00,
   PrivateP00, AccountP00, StratioP00)
P00 <- P00 %>% reduce(full_join, by='CNT')
```

```
P00$CNT <- countrycode(P00$CNT, origin = 'country.name',
   destination = 'iso3c')
SdP00 <- list(SdMP00, SdRP00, SdBookP00, SdImmigP00,
   SdShortP00, SdTeachP00, SdPrivateP00, SdAccountP00,
   SdStratioP00)
SdP00 <- SdP00 %>% reduce(full_join, by='CNT')
SdP00$CNT <- countrycode(SdP00$CNT, origin = 'country.name',
   destination = 'iso3c')
rm(BookP00, ImmigP00, ShortP00, TeachP00, PrivateP00,
   AccountP00, StratioP00,
SdMP00, SdRP00, SdBookP00, SdImmigP00, SdShortP00, SdTeachP00
   , SdPrivateP00, SdAccountP00, SdStratioP00)


#Pisa 2003
pisa03 <- subset(pisa03, select=-c(CNT))
names(pisa03)[names(pisa03) == 'COUNTRY.x'] <- 'CNT'
for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa03[,c(paste("PV",i,"MATH",sep=""),"W_
           FSTUWT")],
        pisa03[,"CNT"], function(x) weighted.mean(x[,1], x
           [,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa03[,c(paste("PV",i,"MATH",sep=""
           ),"W_FSTUWT")],
        pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2])))))
           )}
MeansP03 <- (means1+means2+means3+means4+means5)/5
SdMP03 <- (sd1+sd2+sd3+sd4+sd5)/5
MeansP03 <- as.data.frame(MeansP03)
SdMP03 <- as.data.frame(SdMP03)
MeansP03 <- tibble::rownames_to_column(MeansP03, "CNT")
MeansP03$CNT <- gsub("␣", "", MeansP03$CNT, fixed = TRUE)
MeansP <- merge(MeansP00,MeansP03,by="CNT",all=TRUE)
SdMP03 <- tibble::rownames_to_column(SdMP03, "CNT")
SdMP03$CNT <- gsub("␣", "", SdMP03$CNT, fixed = TRUE)

for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa03[,c(paste("PV",i,"READ",sep=""),"W_
           FSTUWT")],
        pisa03[,"CNT"], function(x) weighted.mean(x[,1], x
           [,2]))))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa03[,c(paste("PV",i,"READ",sep=""
           ),"W_FSTUWT")],
        pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2])))))
           )}
Read03 <- (means1+means2+means3+means4+means5)/5
SdRP03 <- (sd1+sd2+sd3+sd4+sd5)/5
Read03 <- as.data.frame(Read03)
SdRP03 <- as.data.frame(SdRP03)
remove(means1, means2, means3, means4, means5, sd1, sd2, sd3,
    sd4, sd5)
Read03 <- tibble::rownames_to_column(Read03, "CNT")
Read03$CNT <- gsub("␣", "", Read03$CNT, fixed = TRUE)
Read <- merge(Read00,Read03,by="CNT",all=TRUE)
SdRP03 <- tibble::rownames_to_column(SdRP03, "CNT")
SdRP03$CNT <- gsub("␣", "", SdRP03$CNT, fixed = TRUE)

pisa03$ST19Q01 <- dplyr::recode(pisa03$ST19Q01, "0-10␣books"
   =0L, "11-25␣books"=0L,
"26-100␣books"=0L, "101-200␣books"=1L, "201-500␣books"=1L, "
   More␣than␣500␣books"=1L)
BookP03 <- unclass(by(pisa03[,c("ST19Q01","W_FSTUWT")],
   pisa03[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdBookP03 <- sqrt(unclass(by(pisa03[,c("ST19Q01","W_FSTUWT")
   ], pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
BookP03 <- as.data.frame(BookP03)
BookP03 <- tibble::rownames_to_column(BookP03, "CNT")
BookP03$CNT <- gsub("␣", "", BookP03$CNT, fixed = TRUE)
SdBookP03 <- as.data.frame(SdBookP03)
SdBookP03 <- tibble::rownames_to_column(SdBookP03, "CNT")
SdBookP03$CNT <- gsub("␣", "", SdBookP03$CNT, fixed = TRUE)

pisa03$ST15Q01 <- dplyr::recode(pisa03$ST15Q01, "<Country␣of␣
   Test>"=0L, "Other"=1L)
ImmigP03 <- unclass(by(pisa03[,c("ST15Q01","W_FSTUWT")],
   pisa03[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdImmigP03 <- sqrt(unclass(by(pisa03[,c("ST15Q01","W_FSTUWT")
   ], pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
ImmigP03 <- as.data.frame(ImmigP03)
ImmigP03 <- tibble::rownames_to_column(ImmigP03, "CNT")
ImmigP03$CNT <- gsub("␣", "", ImmigP03$CNT, fixed = TRUE)
SdImmigP03 <- as.data.frame(SdImmigP03)
SdImmigP03 <- tibble::rownames_to_column(SdImmigP03, "CNT")
SdImmigP03$CNT <- gsub("␣", "", SdImmigP03$CNT, fixed = TRUE)

pisa03$SC08Q09 <- dplyr::recode(pisa03$SC08Q09, "Not␣at␣all"
   =0L, "Very␣little"=0L, "To␣some␣extent"=1L, "A␣lot"=1L)
ShortP03 <- unclass(by(pisa03[,c("SC08Q09","W_FSTUWT")],
   pisa03[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdShortP03 <- sqrt(unclass(by(pisa03[,c("SC08Q09","W_FSTUWT")
   ], pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
ShortP03 <- as.data.frame(ShortP03)
ShortP03 <- tibble::rownames_to_column(ShortP03, "CNT")
ShortP03$CNT <- gsub("␣", "", ShortP03$CNT, fixed = TRUE)
SdShortP03 <- as.data.frame(SdShortP03)
SdShortP03 <- tibble::rownames_to_column(SdShortP03, "CNT")
SdShortP03$CNT <- gsub("␣", "", SdShortP03$CNT, fixed = TRUE)

pisa03$SC08Q01 <- dplyr::recode(pisa03$SC08Q01, "Not␣at␣all"
   =0L, "Very␣little"=0L, "To␣some␣extent"=1L, "A␣lot"=1L)
TeachP03 <- unclass(by(pisa03[,c("SC08Q01","W_FSTUWT")],
   pisa03[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdTeachP03 <- sqrt(unclass(by(pisa03[,c("SC08Q01","W_FSTUWT")
   ], pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
TeachP03 <- as.data.frame(TeachP03)
TeachP03 <- tibble::rownames_to_column(TeachP03, "CNT")
TeachP03$CNT <- gsub("␣", "", TeachP03$CNT, fixed = TRUE)
SdTeachP03 <- as.data.frame(SdTeachP03)
SdTeachP03 <- tibble::rownames_to_column(SdTeachP03, "CNT")
SdTeachP03$CNT <- gsub("␣", "", SdTeachP03$CNT, fixed = TRUE)

pisa03$SC03Q01 <- dplyr::recode(pisa03$SC03Q01, "Public"=0L,
   "Private"=1L)
```

```
PrivateP03 <- unclass(by(pisa03[,c("SC03Q01","W_FSTUWT")],
   pisa03[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdPrivateP03 <- sqrt(unclass(by(pisa03[,c("SC03Q01","W_FSTUWT
   ")], pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
PrivateP03 <- as.data.frame(PrivateP03)
PrivateP03 <- tibble::rownames_to_column(PrivateP03, "CNT")
PrivateP03$CNT <- gsub("␣", "", PrivateP03$CNT, fixed = TRUE)
SdPrivateP03 <- as.data.frame(SdPrivateP03)
SdPrivateP03 <- tibble::rownames_to_column(SdPrivateP03, "CNT
   ")
SdPrivateP03$CNT <- gsub("␣", "", SdPrivateP03$CNT, fixed =
   TRUE)

pisa03$SC13Q04 <- dplyr::recode(pisa03$SC13Q04, "Yes"=1L, "No
   "=0L)
AccountP03 <- unclass(by(pisa03[,c("SC13Q04","W_FSTUWT")],
   pisa03[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdAccountP03 <- sqrt(unclass(by(pisa03[,c("SC13Q04","W_FSTUWT
   ")], pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
AccountP03 <- as.data.frame(AccountP03)
AccountP03 <- tibble::rownames_to_column(AccountP03, "CNT")
AccountP03$CNT <- gsub("␣", "", AccountP03$CNT, fixed = TRUE)
SdAccountP03 <- as.data.frame(SdAccountP03)
SdAccountP03 <- tibble::rownames_to_column(SdAccountP03, "CNT
   ")
SdAccountP03$CNT <- gsub("␣", "", SdAccountP03$CNT, fixed =
   TRUE)

StratioP03 <- unclass(by(pisa03[,c("STRATIO","W_FSTUWT")],
   pisa03[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdStratioP03 <- sqrt(unclass(by(pisa03[,c("STRATIO","W_FSTUWT
   ")], pisa03[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
StratioP03 <- as.data.frame(StratioP03)
StratioP03 <- tibble::rownames_to_column(StratioP03, "CNT")
StratioP03$CNT <- gsub("␣", "", StratioP03$CNT, fixed = TRUE)
```

```
SdStratioP03 <- as.data.frame(SdStratioP03)
SdStratioP03 <- tibble::rownames_to_column(SdStratioP03, "CNT
   ")
SdStratioP03$CNT <- gsub("␣", "", SdStratioP03$CNT, fixed =
   TRUE)

P03 <- list(MeansP03, BookP03, ImmigP03, ShortP03, TeachP03,
   PrivateP03, AccountP03, StratioP03)
P03 <- P03 %>% reduce(full_join, by='CNT')
P03$CNT <- countrycode(P03$CNT, origin = 'country.name',
   destination = 'iso3c')
SdP03 <- list(SdMP03, SdRP03, SdBookP03, SdImmigP03,
   SdShortP03, SdTeachP03, SdPrivateP03, SdAccountP03,
   SdStratioP03)
SdP03 <- SdP03 %>% reduce(full_join, by='CNT')
SdP03$CNT <- countrycode(SdP03$CNT, origin = 'country.name',
   destination = 'iso3c')
rm(BookP03, ImmigP03, ShortP03, TeachP03, PrivateP03,
   AccountP03, StratioP03,
SdMP03, SdRP03, SdBookP03, SdImmigP03, SdShortP03, SdTeachP03
   , SdPrivateP03, SdAccountP03, SdStratioP03)


#Pisa 2006
names(pisa06)[names(pisa06) == 'COUNTRY'] <- 'CNT'
for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa06[,c(paste("PV",i,"MATH",sep=""),"W_
           FSTUWT")],
        pisa06[,"CNT"], function(x) weighted.mean(x[,1], x
           [,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa06[,c(paste("PV",i,"MATH",sep=""
           ),"W_FSTUWT")],
        pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2])))))
           )}
MeansP06 <- (means1+means2+means3+means4+means5)/5
SdMP06 <- (sd1+sd2+sd3+sd4+sd5)/5
MeansP06 <- as.data.frame(MeansP06)
MeansP06 <- tibble::rownames_to_column(MeansP06, "CNT")
MeansP06$CNT <- gsub("␣", "", MeansP06$CNT, fixed = TRUE)
MeansP <- merge(MeansP,MeansP06,by="CNT",all=TRUE)
```

```
SdMP06 <- as.data.frame(SdMP06)
SdMP06 <- tibble::rownames_to_column(SdMP06, "CNT")
SdMP06$CNT <- gsub("␣", "", SdMP06$CNT, fixed = TRUE)

for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa06[,c(paste("PV",i,"READ",sep=""),"W_
           FSTUWT")],
        pisa06[,"CNT"], function(x) weighted.mean(x[,1], x
           [,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa06[,c(paste("PV",i,"READ",sep=""
           ),"W_FSTUWT")],
        pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2])))))
           )}
Read06 <- (means1+means2+means3+means4+means5)/5
SdRP06 <- (sd1+sd2+sd3+sd4+sd5)/5
Read06 <- as.data.frame(Read06)
remove(means1, means2, means3, means4, means5, sd1, sd2, sd3,
    sd4, sd5)
Read06 <- tibble::rownames_to_column(Read06, "CNT")
Read06$CNT <- gsub("␣", "", Read06$CNT, fixed = TRUE)
#Missing value for USA due to printing error in Reading 2006
Read06[Read06$CNT=="UnitedStates",2] <- NA
Read <- merge(Read,Read06,by="CNT",all=TRUE)
SdRP06 <- as.data.frame(SdRP06)
SdRP06 <- tibble::rownames_to_column(SdRP06, "CNT")
SdRP06$CNT <- gsub("␣", "", SdRP06$CNT, fixed = TRUE)

pisa06$ST15Q01 <- dplyr::recode(pisa06$ST15Q01, "0-10␣books"
   =0L, "11-25␣books"=0L,
"26-100␣books"=0L, "101-200␣books"=1L, "201-500␣books"=1L, "
   More␣than␣500␣books"=1L)
BookP06 <- unclass(by(pisa06[,c("ST15Q01","W_FSTUWT")],
   pisa06[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdBookP06 <- sqrt(unclass(by(pisa06[,c("ST15Q01","W_FSTUWT")
   ], pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
BookP06 <- as.data.frame(BookP06)
BookP06 <- tibble::rownames_to_column(BookP06, "CNT")
BookP06$CNT <- gsub("␣", "", BookP06$CNT, fixed = TRUE)
```

```
SdBookP06 <- as.data.frame(SdBookP06)
SdBookP06 <- tibble::rownames_to_column(SdBookP06, "CNT")
SdBookP06$CNT <- gsub("␣", "", SdBookP06$CNT, fixed = TRUE)

pisa06$ST11Q01 <- dplyr::recode(pisa06$ST11Q01, "Country␣of␣
   test"=0L, "Other␣Country"=1L)
ImmigP06 <- unclass(by(pisa06[,c("ST11Q01","W_FSTUWT")],
   pisa06[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdImmigP06 <- sqrt(unclass(by(pisa06[,c("ST11Q01","W_FSTUWT")
   ], pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
ImmigP06 <- as.data.frame(ImmigP06)
ImmigP06 <- tibble::rownames_to_column(ImmigP06, "CNT")
ImmigP06$CNT <- gsub("␣", "", ImmigP06$CNT, fixed = TRUE)
SdImmigP06 <- as.data.frame(SdImmigP06)
SdImmigP06 <- tibble::rownames_to_column(SdImmigP06, "CNT")
SdImmigP06$CNT <- gsub("␣", "", SdImmigP06$CNT, fixed = TRUE)

pisa06$SC14Q08 <- dplyr::recode(pisa06$SC14Q08, ‘1‘=0L, ‘2‘=0
   L, ‘3‘=1L, ‘4‘=1L)
ShortP06 <- unclass(by(pisa06[,c("SC14Q08","W_FSTUWT")],
   pisa06[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdShortP06 <- sqrt(unclass(by(pisa06[,c("SC14Q08","W_FSTUWT")
   ], pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
ShortP06 <- as.data.frame(ShortP06)
ShortP06 <- tibble::rownames_to_column(ShortP06, "CNT")
ShortP06$CNT <- gsub("␣", "", ShortP06$CNT, fixed = TRUE)
SdShortP06 <- as.data.frame(SdShortP06)
SdShortP06 <- tibble::rownames_to_column(SdShortP06, "CNT")
SdShortP06$CNT <- gsub("␣", "", SdShortP06$CNT, fixed = TRUE)

pisa06$SC14Q02 <- dplyr::recode(pisa06$SC14Q02, ‘1‘=0L, ‘2‘=0
   L, ‘3‘=1L, ‘4‘=1L)
TeachP06 <- unclass(by(pisa06[,c("SC14Q02","W_FSTUWT")],
   pisa06[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdTeachP06 <- sqrt(unclass(by(pisa06[,c("SC14Q02","W_FSTUWT")
   ], pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
TeachP06 <- as.data.frame(TeachP06)
TeachP06 <- tibble::rownames_to_column(TeachP06, "CNT")
TeachP06$CNT <- gsub("␣", "", TeachP06$CNT, fixed = TRUE)
SdTeachP06 <- as.data.frame(SdTeachP06)
SdTeachP06 <- tibble::rownames_to_column(SdTeachP06, "CNT")
SdTeachP06$CNT <- gsub("␣", "", SdTeachP06$CNT, fixed = TRUE)

pisa06$SC02Q01 <- dplyr::recode(pisa06$SC02Q01, ‘1‘=0L, ‘2‘=1
   L)
PrivateP06 <- unclass(by(pisa06[,c("SC02Q01","W_FSTUWT")],
   pisa06[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdPrivateP06 <- sqrt(unclass(by(pisa06[,c("SC02Q01","W_FSTUWT
   ")], pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
PrivateP06 <- as.data.frame(PrivateP06)
PrivateP06 <- tibble::rownames_to_column(PrivateP06, "CNT")
PrivateP06$CNT <- gsub("␣", "", PrivateP06$CNT, fixed = TRUE)
SdPrivateP06 <- as.data.frame(SdPrivateP06)
SdPrivateP06 <- tibble::rownames_to_column(SdPrivateP06, "CNT
   ")
SdPrivateP06$CNT <- gsub("␣", "", SdPrivateP06$CNT, fixed =
   TRUE)

pisa06$SC15Q02 <- dplyr::recode(pisa06$SC15Q02, ‘1‘=1L, ‘2‘=0
   L)
AccountP06 <- unclass(by(pisa06[,c("SC15Q02","W_FSTUWT")],
   pisa06[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdAccountP06 <- sqrt(unclass(by(pisa06[,c("SC15Q02","W_FSTUWT
   ")], pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
AccountP06 <- as.data.frame(AccountP06)
AccountP06 <- tibble::rownames_to_column(AccountP06, "CNT")
AccountP06$CNT <- gsub("␣", "", AccountP06$CNT, fixed = TRUE)
SdAccountP06 <- as.data.frame(SdAccountP06)
SdAccountP06 <- tibble::rownames_to_column(SdAccountP06, "CNT
   ")
SdAccountP06$CNT <- gsub("␣", "", SdAccountP06$CNT, fixed =
   TRUE)
```

```
StratioP06 <- unclass(by(pisa06[,c("STRATIO","W_FSTUWT")],
   pisa06[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdStratioP06 <- sqrt(unclass(by(pisa06[,c("STRATIO","W_FSTUWT
   ")], pisa06[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
StratioP06 <- as.data.frame(StratioP06)
StratioP06 <- tibble::rownames_to_column(StratioP06, "CNT")
StratioP06$CNT <- gsub("␣", "", StratioP06$CNT, fixed = TRUE)
SdStratioP06 <- as.data.frame(SdStratioP06)
SdStratioP06 <- tibble::rownames_to_column(SdStratioP06, "CNT
   ")
SdStratioP06$CNT <- gsub("␣", "", SdStratioP06$CNT, fixed =
   TRUE)

P06 <- list(MeansP06, BookP06, ImmigP06, ShortP06, TeachP06,
   PrivateP06, AccountP06, StratioP06)
P06 <- P06 %>% reduce(full_join, by=’CNT’)
P06$CNT <- countrycode(P06$CNT, origin = ’country.name’,
   destination = ’iso3c’)
SdP06 <- list(SdMP06, SdRP06, SdBookP06, SdImmigP06,
   SdShortP06, SdTeachP06, SdPrivateP06, SdAccountP06,
   SdStratioP06)
SdP06 <- SdP06 %>% reduce(full_join, by=’CNT’)
SdP06$CNT <- countrycode(SdP06$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookP06, ImmigP06, ShortP06, TeachP06, PrivateP06,
   AccountP06, StratioP06,
SdMP06, SdRP06, SdBookP06, SdImmigP06, SdShortP06, SdTeachP06
   , SdPrivateP06, SdAccountP06, SdStratioP06)


#Pisa 2009
for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa09[,c(paste("PV",i,"MATH",sep=""),"W_
           FSTUWT")],
        pisa09[,"CNT"], function(x) weighted.mean(x[,1], x
           [,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa09[,c(paste("PV",i,"MATH",sep=""
            ),"W_FSTUWT")],
        pisa09[,"CNT"], function(x) wtd.var(x[,1], x[,2]))))
            )}
MeansP09 <- (means1+means2+means3+means4+means5)/5
SdMP09 <- (sd1+sd2+sd3+sd4+sd5)/5
MeansP09 <- as.data.frame(MeansP09)
MeansP09 <- tibble::rownames_to_column(MeansP09, "CNT")
MeansP09$CNT <- gsub("␣", "", MeansP09$CNT, fixed = TRUE)
MeansP <- merge(MeansP,MeansP09,by="CNT",all=TRUE)
SdMP09 <- as.data.frame(SdMP09)
SdMP09 <- tibble::rownames_to_column(SdMP09, "CNT")
SdMP09$CNT <- gsub("␣", "", SdMP09$CNT, fixed = TRUE)

for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa09[,c(paste("PV",i,"READ",sep=""),"W_
            FSTUWT")],
        pisa09[,"CNT"], function(x) weighted.mean(x[,1], x
            [,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa09[,c(paste("PV",i,"READ",sep=""
            ),"W_FSTUWT")],
        pisa09[,"CNT"], function(x) wtd.var(x[,1], x[,2]))))
            )}
Read09 <- (means1+means2+means3+means4+means5)/5
SdRP09 <- (sd1+sd2+sd3+sd4+sd5)/5
Read09 <- as.data.frame(Read09)
remove(means1, means2, means3, means4, means5, sd1, sd2, sd3,
    sd4, sd5)
Read09 <- tibble::rownames_to_column(Read09, "CNT")
Read09$CNT <- gsub("␣", "", Read09$CNT, fixed = TRUE)
Read <- merge(Read,Read09,by="CNT",all=TRUE)
SdRP09 <- as.data.frame(SdRP09)
SdRP09 <- tibble::rownames_to_column(SdRP09, "CNT")
SdRP09$CNT <- gsub("␣", "", SdRP09$CNT, fixed = TRUE)

pisa09stu$ST22Q01 <- dplyr::recode(pisa09stu$ST22Q01, "0-10␣
   books"=0L, "11-25␣books"=0L,
"26-100␣books"=0L, "101-200␣books"=1L, "201-500␣books"=1L, "
   More␣than␣500␣books"=1L)
BookP09 <- unclass(by(pisa09stu[,c("ST22Q01","W_FSTUWT")],
   pisa09stu[,"CNT"], function(x) weighted.mean(x[,1], x[,2],
    na.rm=TRUE)))
SdBookP09 <- sqrt(unclass(by(pisa09stu[,c("ST22Q01","W_FSTUWT
   ")], pisa09stu[,"CNT"], function(x) wtd.var(x[,1], x[,2],
   na.rm=TRUE))))
BookP09 <- as.data.frame(BookP09)
BookP09 <- tibble::rownames_to_column(BookP09, "CNT")
BookP09$CNT <- gsub("␣", "", BookP09$CNT, fixed = TRUE)
SdBookP09 <- as.data.frame(SdBookP09)
SdBookP09 <- tibble::rownames_to_column(SdBookP09, "CNT")
SdBookP09$CNT <- gsub("␣", "", SdBookP09$CNT, fixed = TRUE)

pisa09stu$ST17Q01 <- dplyr::recode(pisa09stu$ST17Q01, "
   Country␣of␣test"=0L, "Other␣country"=1L)
ImmigP09 <- unclass(by(pisa09stu[,c("ST17Q01","W_FSTUWT")],
   pisa09stu[,"CNT"], function(x) weighted.mean(x[,1], x[,2],
    na.rm=TRUE)))
SdImmigP09 <- sqrt(unclass(by(pisa09stu[,c("ST17Q01","W_
   FSTUWT")], pisa09stu[,"CNT"], function(x) wtd.var(x[,1], x
   [,2], na.rm=TRUE))))
ImmigP09 <- as.data.frame(ImmigP09)
ImmigP09 <- tibble::rownames_to_column(ImmigP09, "CNT")
ImmigP09$CNT <- gsub("␣", "", ImmigP09$CNT, fixed = TRUE)
SdImmigP09 <- as.data.frame(SdImmigP09)
SdImmigP09 <- tibble::rownames_to_column(SdImmigP09, "CNT")
SdImmigP09$CNT <- gsub("␣", "", SdImmigP09$CNT, fixed = TRUE)

pisa09$SC11Q08 <- dplyr::recode(pisa09$SC11Q08, "Not␣at␣all"
   =0L, "Very␣little"=0L, "To␣some␣extent"=1L, "A␣lot"=1L)
ShortP09 <- unclass(by(pisa09[,c("SC11Q08","W_FSTUWT")],
   pisa09[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdShortP09 <- sqrt(unclass(by(pisa09[,c("SC11Q08","W_FSTUWT")
   ], pisa09[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
ShortP09 <- as.data.frame(ShortP09)
ShortP09 <- tibble::rownames_to_column(ShortP09, "CNT")
ShortP09$CNT <- gsub("␣", "", ShortP09$CNT, fixed = TRUE)
SdShortP09 <- as.data.frame(SdShortP09)
SdShortP09 <- tibble::rownames_to_column(SdShortP09, "CNT")
SdShortP09$CNT <- gsub("␣", "", SdShortP09$CNT, fixed = TRUE)
```

```
pisa09$SC11Q04 <- dplyr::recode(pisa09$SC11Q04, "Not␣at␣all"
   =0L, "Very␣little"=0L, "To␣some␣extent"=1L, "A␣lot"=1L)
TeachP09 <- unclass(by(pisa09[,c("SC11Q04","W_FSTUWT")],
   pisa09[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdTeachP09 <- sqrt(unclass(by(pisa09[,c("SC11Q04","W_FSTUWT")
   ], pisa09[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
TeachP09 <- as.data.frame(TeachP09)
TeachP09 <- tibble::rownames_to_column(TeachP09, "CNT")
TeachP09$CNT <- gsub("␣", "", TeachP09$CNT, fixed = TRUE)
SdTeachP09 <- as.data.frame(SdTeachP09)
SdTeachP09 <- tibble::rownames_to_column(SdTeachP09, "CNT")
SdTeachP09$CNT <- gsub("␣", "", SdTeachP09$CNT, fixed = TRUE)

pisa09$SC02Q01 <- dplyr::recode(pisa09$SC02Q01, "Public"=0L,
   "Private"=1L)
PrivateP09 <- unclass(by(pisa09[,c("SC02Q01","W_FSTUWT")],
   pisa09[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdPrivateP09 <- sqrt(unclass(by(pisa09[,c("SC02Q01","W_FSTUWT
   ")], pisa09[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
PrivateP09 <- as.data.frame(PrivateP09)
PrivateP09 <- tibble::rownames_to_column(PrivateP09, "CNT")
PrivateP09$CNT <- gsub("␣", "", PrivateP09$CNT, fixed = TRUE)
SdPrivateP09 <- as.data.frame(SdPrivateP09)
SdPrivateP09 <- tibble::rownames_to_column(SdPrivateP09, "CNT
   ")
SdPrivateP09$CNT <- gsub("␣", "", SdPrivateP09$CNT, fixed =
   TRUE)

pisa09$SC16Q04 <- dplyr::recode(pisa09$SC16Q04, "Yes"=1L, "No
   "=0L)
AccountP09 <- unclass(by(pisa09[,c("SC16Q04","W_FSTUWT")],
   pisa09[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdAccountP09 <- sqrt(unclass(by(pisa09[,c("SC16Q04","W_FSTUWT
   ")], pisa09[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
AccountP09 <- as.data.frame(AccountP09)
```

```
AccountP09 <- tibble::rownames_to_column(AccountP09, "CNT")
AccountP09$CNT <- gsub("␣", "", AccountP09$CNT, fixed = TRUE)
SdAccountP09 <- as.data.frame(SdAccountP09)
SdAccountP09 <- tibble::rownames_to_column(SdAccountP09, "CNT
   ")
SdAccountP09$CNT <- gsub("␣", "", SdAccountP09$CNT, fixed =
   TRUE)

StratioP09 <- unclass(by(pisa09[,c("STRATIO","W_FSTUWT")],
   pisa09[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdStratioP09 <- sqrt(unclass(by(pisa09[,c("STRATIO","W_FSTUWT
   ")], pisa09[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
StratioP09 <- as.data.frame(StratioP09)
StratioP09 <- tibble::rownames_to_column(StratioP09, "CNT")
StratioP09$CNT <- gsub("␣", "", StratioP09$CNT, fixed = TRUE)
SdStratioP09 <- as.data.frame(SdStratioP09)
SdStratioP09 <- tibble::rownames_to_column(SdStratioP09, "CNT
   ")
SdStratioP09$CNT <- gsub("␣", "", SdStratioP09$CNT, fixed =
   TRUE)

P09 <- list(MeansP09, BookP09, ImmigP09, ShortP09, TeachP09,
   PrivateP09, AccountP09, StratioP09)
P09 <- P09 %>% reduce(full_join, by=’CNT’)
P09$CNT <- countrycode(P09$CNT, origin = ’country.name’,
   destination = ’iso3c’)
SdP09 <- list(SdMP09, SdRP09, SdBookP09, SdImmigP09,
   SdShortP09, SdTeachP09, SdPrivateP09, SdAccountP09,
   SdStratioP09)
SdP09 <- SdP09 %>% reduce(full_join, by=’CNT’)
SdP09$CNT <- countrycode(SdP09$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookP09, ImmigP09, ShortP09, TeachP09, PrivateP09,
   AccountP09, StratioP09,
SdMP09, SdRP09, SdBookP09, SdImmigP09, SdShortP09, SdTeachP09
   , SdPrivateP09, SdAccountP09, SdStratioP09)

rm(pisa09stu, pisa09sch)
```

```
#Pisa 2012
for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa12stu[,c(paste("PV",i,"MATH",sep=""),
           "W_FSTUWT")],
        pisa12stu[,"CNT"], function(x) weighted.mean(x[,1], x
           [,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa12stu[,c(paste("PV",i,"MATH",sep
           =""),"W_FSTUWT")],
        pisa12stu[,"CNT"], function(x) wtd.var(x[,1], x[,2]))
           ))))}
MeansP12 <- (means1+means2+means3+means4+means5)/5
SdMP12 <- (sd1+sd2+sd3+sd4+sd5)/5
MeansP12 <- as.data.frame(MeansP12)
MeansP12 <- tibble::rownames_to_column(MeansP12, "CNT")
MeansP12$CNT <- gsub("␣", "", MeansP12$CNT, fixed = TRUE)
MeansP12$CNT <- dplyr::recode(MeansP12$CNT, "
   UnitedStatesofAmerica"="UnitedStates")
MeansP12 <- MeansP12[-c(51:54),]
MeansP <- merge(MeansP,MeansP12,by="CNT",all=TRUE)
SdMP12 <- as.data.frame(SdMP12)
SdMP12 <- tibble::rownames_to_column(SdMP12, "CNT")
SdMP12$CNT <- gsub("␣", "", SdMP12$CNT, fixed = TRUE)
SdMP12$CNT <- dplyr::recode(SdMP12$CNT, "
   UnitedStatesofAmerica"="UnitedStates")
SdMP12 <- SdMP12[-c(51:54),]

for (i in (1:5)) {assign(paste("means",i, sep=""),
        (unclass(by(pisa12stu[,c(paste("PV",i,"READ",sep=""),
           "W_FSTUWT")],
        pisa12stu[,"CNT"], function(x) weighted.mean(x[,1], x
           [,2])))))}
for (i in (1:5)) {assign(paste("sd",i, sep=""),
        (sqrt(unclass(by(pisa12stu[,c(paste("PV",i,"READ",sep
           =""),"W_FSTUWT")],
        pisa12stu[,"CNT"], function(x) wtd.var(x[,1], x[,2]))
           ))))}
Read12 <- (means1+means2+means3+means4+means5)/5
SdRP12 <- (sd1+sd2+sd3+sd4+sd5)/5
Read12 <- as.data.frame(Read12)
remove(means1, means2, means3, means4, means5, sd1, sd2, sd3,
    sd4, sd5)
Read12 <- tibble::rownames_to_column(Read12, "CNT")
Read12$CNT <- gsub("␣", "", Read12$CNT, fixed = TRUE)
Read12$CNT <- dplyr::recode(Read12$CNT, "
   UnitedStatesofAmerica"="UnitedStates")
Read12 <- Read12[-c(51:54),]
Read <- merge(Read,Read12,by="CNT",all=TRUE)
SdRP12 <- as.data.frame(SdRP12)
SdRP12 <- tibble::rownames_to_column(SdRP12, "CNT")
SdRP12$CNT <- gsub("␣", "", SdRP12$CNT, fixed = TRUE)
SdRP12$CNT <- dplyr::recode(SdRP12$CNT, "
   UnitedStatesofAmerica"="UnitedStates")
SdRP12 <- SdRP12[-c(51:54),]

pisa12stu$ST28Q01 <- dplyr::recode(pisa12stu$ST28Q01, "0-10␣
   books␣"=0L, "11-25␣books␣"=0L,
"26-100␣books␣"=0L, "101-200␣books␣"=1L, "201-500␣books␣"=1L,
    "More␣than␣500␣books"=1L)
BookP12 <- unclass(by(pisa12stu[,c("ST28Q01","W_FSTUWT")],
   pisa12stu[,"CNT"], function(x) weighted.mean(x[,1], x[,2],
    na.rm=TRUE)))
SdBookP12 <- sqrt(unclass(by(pisa12stu[,c("ST28Q01","W_FSTUWT
   ")], pisa12stu[,"CNT"], function(x) wtd.var(x[,1], x[,2],
   na.rm=TRUE))))
BookP12 <- as.data.frame(BookP12)
BookP12 <- tibble::rownames_to_column(BookP12, "CNT")
BookP12$CNT <- gsub("␣", "", BookP12$CNT, fixed = TRUE)
SdBookP12 <- as.data.frame(SdBookP12)
SdBookP12 <- tibble::rownames_to_column(SdBookP12, "CNT")
SdBookP12$CNT <- gsub("␣", "", SdBookP12$CNT, fixed = TRUE)

pisa12stu$ST20Q01 <- dplyr::recode(pisa12stu$ST20Q01, "
   Country␣of␣test"=0L, "Other␣country"=1L)
ImmigP12 <- unclass(by(pisa12stu[,c("ST20Q01","W_FSTUWT")],
   pisa12stu[,"CNT"], function(x) weighted.mean(x[,1], x[,2],
    na.rm=TRUE)))
SdImmigP12 <- sqrt(unclass(by(pisa12stu[,c("ST20Q01","W_
   FSTUWT")], pisa12stu[,"CNT"], function(x) wtd.var(x[,1], x
   [,2], na.rm=TRUE))))
ImmigP12 <- as.data.frame(ImmigP12)
ImmigP12 <- tibble::rownames_to_column(ImmigP12, "CNT")
```

```
ImmigP12$CNT <- gsub("␣", "", ImmigP12$CNT, fixed = TRUE)
SdImmigP12 <- as.data.frame(SdImmigP12)
SdImmigP12 <- tibble::rownames_to_column(SdImmigP12, "CNT")
SdImmigP12$CNT <- gsub("␣", "", SdImmigP12$CNT, fixed = TRUE)

pisa12$SC14Q06 <- dplyr::recode(pisa12$SC14Q06, "Not␣at␣all"
   =0L, "Very␣little"=0L, "To␣some␣extent"=1L, "A␣lot"=1L)
ShortP12 <- unclass(by(pisa12[,c("SC14Q06","W_FSTUWT")],
   pisa12[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdShortP12 <- sqrt(unclass(by(pisa12[,c("SC14Q06","W_FSTUWT")
   ], pisa12[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
ShortP12 <- as.data.frame(ShortP12)
ShortP12 <- tibble::rownames_to_column(ShortP12, "CNT")
ShortP12$CNT <- gsub("␣", "", ShortP12$CNT, fixed = TRUE)
SdShortP12 <- as.data.frame(SdShortP12)
SdShortP12 <- tibble::rownames_to_column(SdShortP12, "CNT")
SdShortP12$CNT <- gsub("␣", "", SdShortP12$CNT, fixed = TRUE)

pisa12$SC14Q02 <- dplyr::recode(pisa12$SC14Q02, "Not␣at␣all"
   =0L, "Very␣little"=0L, "To␣some␣extent"=1L, "A␣lot"=1L)
TeachP12 <- unclass(by(pisa12[,c("SC14Q02","W_FSTUWT")],
   pisa12[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdTeachP12 <- sqrt(unclass(by(pisa12[,c("SC14Q02","W_FSTUWT")
   ], pisa12[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.rm
   =TRUE))))
TeachP12 <- as.data.frame(TeachP12)
TeachP12 <- tibble::rownames_to_column(TeachP12, "CNT")
TeachP12$CNT <- gsub("␣", "", TeachP12$CNT, fixed = TRUE)
SdTeachP12 <- as.data.frame(SdTeachP12)
SdTeachP12 <- tibble::rownames_to_column(SdTeachP12, "CNT")
SdTeachP12$CNT <- gsub("␣", "", SdTeachP12$CNT, fixed = TRUE)

pisa12$SC01Q01 <- dplyr::recode(pisa12$SC01Q01, "Public"=0L,
   "Private"=1L)
PrivateP12 <- unclass(by(pisa12[,c("SC01Q01","W_FSTUWT")],
   pisa12[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdPrivateP12 <- sqrt(unclass(by(pisa12[,c("SC01Q01","W_FSTUWT
   ")], pisa12[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
PrivateP12 <- as.data.frame(PrivateP12)
PrivateP12 <- tibble::rownames_to_column(PrivateP12, "CNT")
PrivateP12$CNT <- gsub("␣", "", PrivateP12$CNT, fixed = TRUE)
SdPrivateP12 <- as.data.frame(SdPrivateP12)
SdPrivateP12 <- tibble::rownames_to_column(SdPrivateP12, "CNT
   ")
SdPrivateP12$CNT <- gsub("␣", "", SdPrivateP12$CNT, fixed =
   TRUE)

pisa12$SC18Q04 <- dplyr::recode(pisa12$SC18Q04, "Yes"=1L, "No
   "=0L)
AccountP12 <- unclass(by(pisa12[,c("SC18Q04","W_FSTUWT")],
   pisa12[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdAccountP12 <- sqrt(unclass(by(pisa12[,c("SC18Q04","W_FSTUWT
   ")], pisa12[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
AccountP12 <- as.data.frame(AccountP12)
AccountP12 <- tibble::rownames_to_column(AccountP12, "CNT")
AccountP12$CNT <- gsub("␣", "", AccountP12$CNT, fixed = TRUE)
SdAccountP12 <- as.data.frame(SdAccountP12)
SdAccountP12 <- tibble::rownames_to_column(SdAccountP12, "CNT
   ")
SdAccountP12$CNT <- gsub("␣", "", SdAccountP12$CNT, fixed =
   TRUE)

StratioP12 <- unclass(by(pisa12[,c("STRATIO","W_FSTUWT")],
   pisa12[,"CNT"], function(x) weighted.mean(x[,1], x[,2], na
   .rm=TRUE)))
SdStratioP12 <- sqrt(unclass(by(pisa12[,c("STRATIO","W_FSTUWT
   ")], pisa12[,"CNT"], function(x) wtd.var(x[,1], x[,2], na.
   rm=TRUE))))
StratioP12 <- as.data.frame(StratioP12)
StratioP12 <- tibble::rownames_to_column(StratioP12, "CNT")
StratioP12$CNT <- gsub("␣", "", StratioP12$CNT, fixed = TRUE)
SdStratioP12 <- as.data.frame(SdStratioP12)
SdStratioP12 <- tibble::rownames_to_column(SdStratioP12, "CNT
   ")
SdStratioP12$CNT <- gsub("␣", "", SdStratioP12$CNT, fixed =
   TRUE)

P12 <- list(BookP12, ImmigP12, ShortP12, TeachP12, PrivateP12
   , AccountP12, StratioP12)
P12 <- P12 %>% reduce(full_join, by=’CNT’)
P12$CNT <- dplyr::recode(P12$CNT, "UnitedStatesofAmerica"="
   UnitedStates")
P12 <- P12[-c(51:54),]
P12 <- merge(MeansP12,P12,by="CNT",all=TRUE)
P12$CNT <- countrycode(P12$CNT, origin = ’country.name’,
   destination = ’iso3c’)
SdP12 <- list(SdBookP12, SdImmigP12, SdShortP12, SdTeachP12,
   SdPrivateP12, SdAccountP12, SdStratioP12)
SdP12 <- SdP12 %>% reduce(full_join, by=’CNT’)
SdP12$CNT <- dplyr::recode(SdP12$CNT, "UnitedStatesofAmerica"
   ="UnitedStates")
SdP12 <- SdP12[-c(51:54),]
SdP12 <- merge(SdRP12, SdP12, by="CNT", all=TRUE)
SdP12 <- merge(SdMP12, SdP12, by="CNT", all=TRUE)
SdP12$CNT <- countrycode(SdP12$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookP12, ImmigP12, ShortP12, TeachP12, PrivateP12,
   AccountP12, StratioP12,
SdMP12, SdRP12, SdBookP12, SdImmigP12, SdShortP12, SdTeachP12
   , SdPrivateP12, SdAccountP12, SdStratioP12)

rm(pisa12stu, pisa12sch)

MeansP$CNT <- countrycode(MeansP$CNT, origin = ’country.name’
   , destination = ’iso3c’)
Read$CNT <- countrycode(Read$CNT, origin = ’country.name’,
   destination = ’iso3c’)


#Pisa 2015
MeansP15 <- pisa.mean.pv(pvlabel = "MATH", by = "CNT", data =
    pisa15)
colnames(MeansP15)[3] <- "MeansP15"
SdMP15 <- as.data.frame(MeansP15[c(1,5)])
colnames(SdMP15)[2] <- "SdMP15"
MeansP15 <- MeansP15[c(1,3)]
```

```
MeansP <- merge(MeansP,MeansP15,by="CNT",all=TRUE)

Read15 <- pisa.mean.pv(pvlabel = "READ", by = "CNT", data =
   pisa15)
colnames(Read15)[3] <- "Read15"
SdRP15 <- as.data.frame(Read15[c(1,5)])
colnames(SdRP15)[2] <- "SdRP15"
Read15 <- Read15[c(1,3)]
Read <- merge(Read,Read15,by="CNT",all=TRUE)

pisa15$ST013Q01TA <- dplyr::recode(pisa15$ST013Q01TA, ‘1‘=0L,
    ‘2‘=0L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L, ‘6‘=1L)
BookP15 <- pisa.mean(variable=’ST013Q01TA’, by=’CNT’, data=
   pisa15)
colnames(BookP15)[3] <- "BookP15"
SdBookP15 <- as.data.frame(BookP15[c(1,5)])
colnames(SdBookP15)[2] <- "SdBookP15"
BookP15 <- BookP15[c(1,3)]

pisa15$ST019AQ01T <- dplyr::recode(pisa15$ST019AQ01T, ‘1‘=0L,
    ‘2‘=1L)
ImmigP15 <- pisa.mean(variable=’ST019AQ01T’, by=’CNT’, data=
   pisa15)
colnames(ImmigP15)[3] <- "ImmigP15"
SdImmigP15 <- as.data.frame(ImmigP15[c(1,5)])
colnames(SdImmigP15)[2] <- "SdImmigP15"
ImmigP15 <- ImmigP15[c(1,3)]

pisa15$SC017Q05NA <- dplyr::recode(pisa15$SC017Q05NA, ‘1‘=0L,
    ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortP15 <- pisa.mean(variable=’SC017Q05NA’, by=’CNT’, data=
   pisa15)
colnames(ShortP15)[3] <- "ShortP15"
SdShortP15 <- as.data.frame(ShortP15[c(1,5)])
colnames(SdShortP15)[2] <- "SdShortP15"
ShortP15 <- ShortP15[c(1,3)]

pisa15$SC017Q01NA <- dplyr::recode(pisa15$SC017Q01NA, ‘1‘=0L,
    ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachP15 <- pisa.mean(variable=’SC017Q01NA’, by=’CNT’, data=
   pisa15)
```

```
colnames(TeachP15)[3] <- "TeachP15"
SdTeachP15 <- as.data.frame(TeachP15[c(1,5)])
colnames(SdTeachP15)[2] <- "SdTeachP15"
TeachP15 <- TeachP15[c(1,3)]

pisa15$SC013Q01TA <- dplyr::recode(pisa15$SC013Q01TA, ‘1‘=0L,
    ‘2‘=1L)
PrivateP15 <- pisa.mean(variable=’SC013Q01TA’, by=’CNT’, data
   =pisa15)
colnames(PrivateP15)[3] <- "PrivateP15"
SdPrivateP15 <- as.data.frame(PrivateP15[c(1,5)])
colnames(SdPrivateP15)[2] <- "SdPrivateP15"
PrivateP15 <- PrivateP15[c(1,3)]

pisa15$SC035Q05TA <- dplyr::recode(pisa15$SC035Q05TA, ‘1‘=1L,
    ‘2‘=0L)
AccountP15 <- pisa.mean(variable=’SC035Q05TA’, by=’CNT’, data
   =pisa15)
colnames(AccountP15)[3] <- "AccountP15"
SdAccountP15 <- as.data.frame(AccountP15[c(1,5)])
colnames(SdAccountP15)[2] <- "SdAccountP15"
AccountP15 <- AccountP15[c(1,3)]

StratioP15 <- pisa.mean(variable=’STRATIO’, by=’CNT’, data=
   pisa15)
colnames(StratioP15)[3] <- "StratioP15"
SdStratioP15 <- as.data.frame(StratioP15[c(1,5)])
colnames(SdStratioP15)[2] <- "SdStratioP15"
StratioP15 <- StratioP15[c(1,3)]

P15 <- list(MeansP15, BookP15, ImmigP15, ShortP15, TeachP15,
   PrivateP15, AccountP15, StratioP15)
P15 <- P15 %>% reduce(full_join, by=’CNT’)
SdP15 <- list(SdMP15, SdRP15, SdBookP15, SdImmigP15,
   SdShortP15, SdTeachP15, SdPrivateP15, SdAccountP15,
   SdStratioP15)
SdP15 <- SdP15 %>% reduce(full_join, by=’CNT’)
rm(BookP15, ImmigP15, ShortP15, TeachP15, PrivateP15,
   AccountP15, StratioP15,
SdMP15, SdRP15, SdBookP15, SdImmigP15, SdShortP15, SdTeachP15
   , SdPrivateP15, SdAccountP15, SdStratioP15)
```

```
#Pisa 2018
MeansP18 <- pisa.mean.pv(pvlabel = "MATH", by = "CNT", data =
    pisa18)
colnames(MeansP18)[3] <- "MeansP18"
MeansP18 <- MeansP18[c(1,3)]
MeansP <- merge(MeansP,MeansP18,by="CNT",all=TRUE)
#MeansP <- MeansP[, -grep("COUNTRY*", colnames(MeansP))]

Read18 <- pisa.mean.pv(pvlabel = "READ", by = "CNT", data =
   pisa18)
colnames(Read18)[3] <- "Read18"
Read18 <- Read18[c(1,3)]
Read <- merge(Read,Read18,by="CNT",all=TRUE)
#MeansP <- MeansP[, -grep("COUNTRY*", colnames(MeansP))]

pisa18$ST013Q01TA <- dplyr::recode(pisa18$ST013Q01TA, ‘1‘=0L,
    ‘2‘=0L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L, ‘6‘=1L)
BookP18 <- pisa.mean(variable=’ST013Q01TA’, by=’CNT’, data=
   pisa18)
colnames(BookP18)[3] <- "BookP18"
BookP18 <- BookP18[c(1,3)]

pisa18$ST019AQ01T <- dplyr::recode(pisa18$ST019AQ01T, ‘1‘=0L,
    ‘2‘=1L)
ImmigP18 <- pisa.mean(variable=’ST019AQ01T’, by=’CNT’, data=
   pisa18)
colnames(ImmigP18)[3] <- "ImmigP18"
ImmigP18 <- ImmigP18[c(1,3)]

pisa18$SC017Q05NA <- dplyr::recode(pisa18$SC017Q05NA, ‘1‘=0L,
    ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortP18 <- pisa.mean(variable=’SC017Q05NA’, by=’CNT’, data=
   pisa18)
colnames(ShortP18)[3] <- "ShortP18"
ShortP18 <- ShortP18[c(1,3)]

pisa18$SC017Q01NA <- dplyr::recode(pisa18$SC017Q01NA, ‘1‘=0L,
    ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachP18 <- pisa.mean(variable=’SC017Q01NA’, by=’CNT’, data=
   pisa18)
colnames(TeachP18)[3] <- "TeachP18"
TeachP18 <- TeachP18[c(1,3)]

pisa18$SC013Q01TA <- dplyr::recode(pisa18$SC013Q01TA, ‘1‘=0L,
    ‘2‘=1L)
PrivateP18 <- pisa.mean(variable=’SC013Q01TA’, by=’CNT’, data
   =pisa18)
colnames(PrivateP18)[3] <- "PrivateP18"
PrivateP18 <- PrivateP18[c(1,3)]

pisa18$SC154Q05WA <- dplyr::recode(pisa18$SC154Q05WA, ‘1‘=1L,
    ‘2‘=0L)
AccountP18 <- pisa.mean(variable=’SC154Q05WA’, by=’CNT’, data
   =pisa18)
colnames(AccountP18)[3] <- "AccountP18"
AccountP18 <- AccountP18[c(1,3)]

StratioP18 <- pisa.mean(variable=’STRATIO’, by=’CNT’, data=
   pisa18)
colnames(StratioP18)[3] <- "StratioP18"
StratioP18 <- StratioP18[c(1,3)]

P18 <- list(MeansP18, BookP18, ImmigP18, ShortP18, TeachP18,
   PrivateP18, AccountP18, StratioP18)
P18 <- P18 %>% reduce(full_join, by=’CNT’)
rm(BookP18, ImmigP18, ShortP18, TeachP18, PrivateP18,
   AccountP18, StratioP18)

rm(MeansP00, MeansP03, MeansP06, MeansP09, MeansP12, MeansP15
   , MeansP18)
rm(Read00, Read03, Read06, Read09, Read12, Read15, Read18)

#Selection of countries
MeansP <- MeansP[MeansP$CNT %in% list("AUS","AUT","BEL","BGR"
   ,"CAN","CHE","CZE",
"DEU","DNK","ESP","FRA","FIN","GBR","IRL",
"ISL","ISR","ITA","LUX","LVA","NLD","NOR",
"NZL","RUS","SWE","USA"), ]

Read <- Read[Read$CNT %in% list("AUS","AUT","BEL","BGR","CAN"
   ,"CHE","CZE",
"DEU","DNK","ESP","FRA","FIN","GBR","IRL",
"ISL","ISR","ITA","LUX","LVA","NLD","NOR",
"NZL","RUS","SWE","USA"), ]


#Timss 1999
MeansT99 <- timss.mean.pv(pvlabel="BSMMAT", by= c("IDCNTRYL")
   , data=timss99)
colnames(MeansT99)[3] <- "MeansT99"
SdMT99 <- MeansT99[c(1,5)]
colnames(SdMT99)[2] <- "SdMT99"
names(SdMT99)[names(SdMT99) == ’IDCNTRYL’] <- ’CNT’
SdMT99$CNT <- gsub("␣", "", SdMT99$CNT, fixed = TRUE)
MeansT99 <- MeansT99[c(1,3)]

timss99$BSBGBOOK <- dplyr::recode(timss99$BSBGBOOK, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L)
BookT99 <- timss.mean(variable=’BSBGBOOK’, by=’IDCNTRYL’,
   data=timss99)
colnames(BookT99)[3] <- "BookT99"
BookT99 <- BookT99[c(1,3)]
SdBookT99 <- as.data.frame(sqrt(unclass(by(timss99[,c("
   BSBGBOOK","TOTWGT")], timss99[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdBookT99 <- tibble::rownames_to_column(SdBookT99, "CNT")
SdBookT99$CNT <- gsub("␣", "", SdBookT99$CNT, fixed = TRUE)
colnames(SdBookT99)[2] <- "SdBookT99"

timss99$BSBGBRN1 <- dplyr::recode(timss99$BSBGBRN1, ‘1‘=0L,
   ‘2‘=1L)
ImmigT99 <- timss.mean(variable=’BSBGBRN1’, by=’IDCNTRYL’,
   data=timss99)
colnames(ImmigT99)[3] <- "ImmigT99"
ImmigT99 <- ImmigT99[c(1,3)]
SdImmigT99 <- as.data.frame(sqrt(unclass(by(timss99[,c("
   BSBGBRN1","TOTWGT")], timss99[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdImmigT99 <- tibble::rownames_to_column(SdImmigT99, "CNT")
SdImmigT99$CNT <- gsub("␣", "", SdImmigT99$CNT, fixed = TRUE)
colnames(SdImmigT99)[2] <- "SdImmigT99"
```

```
timss99$BCBGST01 <- dplyr::recode(timss99$BCBGST01, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortT99 <- timss.mean(variable=’BCBGST01’, by=’IDCNTRYL’,
   data=timss99)
colnames(ShortT99)[3] <- "ShortT99"
ShortT99 <- ShortT99[c(1,3)]
SdShortT99 <- as.data.frame(sqrt(unclass(by(timss99[,c("
   BCBGST01","TOTWGT")], timss99[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdShortT99 <- tibble::rownames_to_column(SdShortT99, "CNT")
SdShortT99$CNT <- gsub("␣", "", SdShortT99$CNT, fixed = TRUE)
colnames(SdShortT99)[2] <- "SdShortT99"

timss99$BCBSST18 <- dplyr::recode(timss99$BCBSST18, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachT99 <- timss.mean(variable=’BCBSST18’, by=’IDCNTRYL’,
   data=timss99)
colnames(TeachT99)[3] <- "TeachT99"
TeachT99 <- TeachT99[c(1,3)]
SdTeachT99 <- as.data.frame(sqrt(unclass(by(timss99[,c("
   BCBSST18","TOTWGT")], timss99[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdTeachT99 <- tibble::rownames_to_column(SdTeachT99, "CNT")
SdTeachT99$CNT <- gsub("␣", "", SdTeachT99$CNT, fixed = TRUE)
colnames(SdTeachT99)[2] <- "SdTeachT99"

T99 <- list(MeansT99, BookT99, ImmigT99, ShortT99, TeachT99)
T99 <- T99 %>% reduce(full_join, by=’IDCNTRYL’)
names(T99)[names(T99) == ’IDCNTRYL’] <- ’CNT’
T99$CNT <- dplyr::recode(T99$CNT, "England"="UnitedKingdom")
T99$CNT <- countrycode(T99$CNT, origin = ’country.name’,
   destination = ’iso3c’)
SdT99 <- list(SdMT99, SdBookT99, SdImmigT99, SdShortT99,
   SdTeachT99)
SdT99 <- SdT99 %>% reduce(full_join, by=’CNT’)
SdT99$CNT <- dplyr::recode(SdT99$CNT, "England"="
   UnitedKingdom")
SdT99$CNT <- countrycode(SdT99$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookT99, ImmigT99, ShortT99, TeachT99, SdMT99, SdBookT99,
   SdImmigT99, SdShortT99, SdTeachT99)

#Timss 2003
MeansT03 <- timss.mean.pv(pvlabel="BSMMAT", by= c("IDCNTRYL")
   , data=timss03)
colnames(MeansT03)[3] <- "MeansT03"
MeansT03 <- MeansT03[c(1,3)]
MeansT03 <- MeansT03[is.na(MeansT03$IDCNTRYL)==0,]
MeansT <- merge(MeansT99,MeansT03,by="IDCNTRYL",all=TRUE)

timss03$BSBGBOOK <- dplyr::recode(timss03$BSBGBOOK, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L)
BookT03 <- timss.mean(variable=’BSBGBOOK’, by=’IDCNTRYL’,
   data=timss03)
colnames(BookT03)[3] <- "BookT03"
BookT03 <- BookT03[c(1,3)]

timss03$BSBGBORN <- dplyr::recode(timss03$BSBGBORN, ‘1‘=0L,
   ‘2‘=1L)
ImmigT03 <- timss.mean(variable=’BSBGBORN’, by=’IDCNTRYL’,
   data=timss03)
colnames(ImmigT03)[3] <- "ImmigT03"
ImmigT03 <- ImmigT03[c(1,3)]

timss03$BCBGST01 <- dplyr::recode(timss03$BCBGST01, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortT03 <- timss.mean(variable=’BCBGST01’, by=’IDCNTRYL’,
   data=timss03)
colnames(ShortT03)[3] <- "ShortT03"
ShortT03 <- ShortT03[c(1,3)]

timss03$BCBGSH18 <- dplyr::recode(timss03$BCBGSH18, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachT03 <- timss.mean(variable=’BCBGSH18’, by=’IDCNTRYL’,
   data=timss03)
colnames(TeachT03)[3] <- "TeachT03"
TeachT03 <- TeachT03[c(1,3)]

T03 <- list(MeansT03, BookT03, ImmigT03, ShortT03, TeachT03)
T03 <- T03 %>% reduce(full_join, by=’IDCNTRYL’)
names(T03)[names(T03) == ’IDCNTRYL’] <- ’CNT’
```

```
T03$CNT <- dplyr::recode(T03$CNT, "England"="UnitedKingdom")
T03$CNT <- countrycode(T03$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookT03, ImmigT03, ShortT03, TeachT03)

#Timss 2007
MeansT07 <- timss.mean.pv(pvlabel="BSMMAT", by= c("IDCNTRYL")
   , data=timss07)
colnames(MeansT07)[3] <- "MeansT07"
SdMT07 <- MeansT07[c(1,5)]
colnames(SdMT07)[2] <- "SdMT07"
names(SdMT07)[names(SdMT07) == ’IDCNTRYL’] <- ’CNT’
SdMT07$CNT <- gsub("␣", "", SdMT07$CNT, fixed = TRUE)
MeansT07 <- MeansT07[c(1,3)]
MeansT <- merge(MeansT,MeansT07,by="IDCNTRYL",all=TRUE)

timss07$BS4GBOOK <- dplyr::recode(timss07$BS4GBOOK, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L)
BookT07 <- timss.mean(variable=’BS4GBOOK’, by=’IDCNTRYL’,
   data=timss07)
colnames(BookT07)[3] <- "BookT07"
BookT07 <- BookT07[c(1,3)]
SdBookT07 <- as.data.frame(sqrt(unclass(by(timss07[,c("
   BS4GBOOK","TOTWGT")], timss07[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdBookT07 <- tibble::rownames_to_column(SdBookT07, "CNT")
SdBookT07$CNT <- gsub("␣", "", SdBookT07$CNT, fixed = TRUE)
colnames(SdBookT07)[2] <- "SdBookT07"

timss07$BS4GBORN <- dplyr::recode(timss07$BS4GBORN, ‘1‘=0L,
   ‘2‘=1L)
ImmigT07 <- timss.mean(variable=’BS4GBORN’, by=’IDCNTRYL’,
   data=timss07)
colnames(ImmigT07)[3] <- "ImmigT07"
ImmigT07 <- ImmigT07[c(1,3)]
SdImmigT07 <- as.data.frame(sqrt(unclass(by(timss07[,c("
   BS4GBORN","TOTWGT")], timss07[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdImmigT07 <- tibble::rownames_to_column(SdImmigT07, "CNT")
SdImmigT07$CNT <- gsub("␣", "", SdImmigT07$CNT, fixed = TRUE)
colnames(SdImmigT07)[2] <- "SdImmigT07"
```

```
timss07$BC4GST01 <- dplyr::recode(timss07$BC4GST01, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortT07 <- timss.mean(variable=’BC4GST01’, by=’IDCNTRYL’,
   data=timss07)
colnames(ShortT07)[3] <- "ShortT07"
ShortT07 <- ShortT07[c(1,3)]
SdShortT07 <- as.data.frame(sqrt(unclass(by(timss07[,c("
   BC4GST01","TOTWGT")], timss07[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdShortT07 <- tibble::rownames_to_column(SdShortT07, "CNT")
SdShortT07$CNT <- gsub("␣", "", SdShortT07$CNT, fixed = TRUE)
colnames(SdShortT07)[2] <- "SdShortT07"

timss07$BC4GSH18 <- dplyr::recode(timss07$BC4GSH18, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachT07 <- timss.mean(variable=’BC4GSH18’, by=’IDCNTRYL’,
   data=timss07)
colnames(TeachT07)[3] <- "TeachT07"
TeachT07 <- TeachT07[c(1,3)]
SdTeachT07 <- as.data.frame(sqrt(unclass(by(timss07[,c("
   BC4GSH18","TOTWGT")], timss07[,"IDCNTRYL"], function(x)
   wtd.var(x[,1], x[,2], na.rm=TRUE)))))
SdTeachT07 <- tibble::rownames_to_column(SdTeachT07, "CNT")
SdTeachT07$CNT <- gsub("␣", "", SdTeachT07$CNT, fixed = TRUE)
colnames(SdTeachT07)[2] <- "SdTeachT07"

T07 <- list(MeansT07, BookT07, ImmigT07, ShortT07, TeachT07)
T07 <- T07 %>% reduce(full_join, by=’IDCNTRYL’)
names(T07)[names(T07) == ’IDCNTRYL’] <- ’CNT’
T07$CNT <- dplyr::recode(T07$CNT, "England"="UnitedKingdom")
T07$CNT <- countrycode(T07$CNT, origin = ’country.name’,
   destination = ’iso3c’)
SdT07 <- list(SdMT07, SdBookT07, SdImmigT07, SdShortT07,
   SdTeachT07)
SdT07 <- SdT07 %>% reduce(full_join, by=’CNT’)
SdT07$CNT <- dplyr::recode(SdT07$CNT, "England"="
   UnitedKingdom")
SdT07$CNT <- countrycode(SdT07$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookT07, ImmigT07, ShortT07, TeachT07, SdMT07, SdBookT07,
   SdImmigT07, SdShortT07, SdTeachT07)

#Timss 2011
MeansT11 <- timss.mean.pv(pvlabel="BSMMAT", by= c("IDCNTRYL")
   , data=timss11)
colnames(MeansT11)[3] <- "MeansT11"
MeansT11 <- MeansT11[c(1,3)]
MeansT <- merge(MeansT,MeansT11,by="IDCNTRYL",all=TRUE)

timss11$BSBG04 <- dplyr::recode(timss11$BSBG04, ‘1‘=0L, ‘2‘=0
   L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L)
BookT11 <- timss.mean(variable=’BSBG04’, by=’IDCNTRYL’, data=
   timss11)
colnames(BookT11)[3] <- "BookT11"
BookT11 <- BookT11[c(1,3)]

timss11$BSBG09A <- dplyr::recode(timss11$BSBG09A, ‘1‘=0L,
   ‘2‘=1L)
ImmigT11 <- timss.mean(variable=’BSBG09A’, by=’IDCNTRYL’,
   data=timss11)
colnames(ImmigT11)[3] <- "ImmigT11"
ImmigT11 <- ImmigT11[c(1,3)]

timss11$BCBG09AA <- dplyr::recode(timss11$BCBG09AA, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortT11 <- timss.mean(variable=’BCBG09AA’, by=’IDCNTRYL’,
   data=timss11)
colnames(ShortT11)[3] <- "ShortT11"
ShortT11 <- ShortT11[c(1,3)]

timss11$BCBG09B <- dplyr::recode(timss11$BCBG09B, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachT11 <- timss.mean(variable=’BCBG09B’, by=’IDCNTRYL’,
   data=timss11)
colnames(TeachT11)[3] <- "TeachT11"
TeachT11 <- TeachT11[c(1,3)]

T11 <- list(MeansT11, BookT11, ImmigT11, ShortT11, TeachT11)
T11 <- T11 %>% reduce(full_join, by=’IDCNTRYL’)
names(T11)[names(T11) == ’IDCNTRYL’] <- ’CNT’
T11$CNT <- dplyr::recode(T11$CNT, "England"="UnitedKingdom")
```

```
T11$CNT <- countrycode(T11$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookT11, ImmigT11, ShortT11, TeachT11)

#Timss 2015
MeansT15 <- timss.mean.pv(pvlabel="BSMMAT", by= c("IDCNTRYL")
   , data=timss15)
colnames(MeansT15)[3] <- "MeansT15"
MeansT15 <- MeansT15[c(1,3)]
MeansT <- merge(MeansT,MeansT15,by="IDCNTRYL",all=TRUE)

timss15$BSBG04 <- dplyr::recode(timss15$BSBG04, ‘1‘=0L, ‘2‘=0
   L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L)
BookT15 <- timss.mean(variable=’BSBG04’, by=’IDCNTRYL’, data=
   timss15)
colnames(BookT15)[3] <- "BookT15"
BookT15 <- BookT15[c(1,3)]

timss15$BSBG10A <- dplyr::recode(timss15$BSBG10A, ‘1‘=0L,
   ‘2‘=1L)
ImmigT15 <- timss.mean(variable=’BSBG10A’, by=’IDCNTRYL’,
   data=timss15)
colnames(ImmigT15)[3] <- "ImmigT15"
ImmigT15 <- ImmigT15[c(1,3)]

timss15$BCBG13AA <- dplyr::recode(timss15$BCBG13AA, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortT15 <- timss.mean(variable=’BCBG13AA’, by=’IDCNTRYL’,
   data=timss15)
colnames(ShortT15)[3] <- "ShortT15"
ShortT15 <- ShortT15[c(1,3)]

timss15$BCBG13BA <- dplyr::recode(timss15$BCBG13BA, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachT15 <- timss.mean(variable=’BCBG13BA’, by=’IDCNTRYL’,
   data=timss15)
colnames(TeachT15)[3] <- "TeachT15"
TeachT15 <- TeachT15[c(1,3)]

T15 <- list(MeansT15, BookT15, ImmigT15, ShortT15, TeachT15)
T15 <- T15 %>% reduce(full_join, by=’IDCNTRYL’)
```

```
names(T15)[names(T15) == ’IDCNTRYL’] <- ’CNT’
T15$CNT <- dplyr::recode(T15$CNT, "England"="UnitedKingdom")
T15$CNT <- countrycode(T15$CNT, origin = ’country.name’,
   destination = ’iso3c’)
rm(BookT15, ImmigT15, ShortT15, TeachT15)

#Timss 2019
MeansT19 <- timss.mean.pv(pvlabel="BSMMAT", by= c("IDCNTRYL")
   , data=timss19)
colnames(MeansT19)[3] <- "MeansT19"
MeansT19 <- MeansT19[c(1,3)]
MeansT <- merge(MeansT,MeansT19,by="IDCNTRYL",all=TRUE)

timss19$BSBG04 <- dplyr::recode(timss19$BSBG04, ‘1‘=0L, ‘2‘=0
   L, ‘3‘=0L, ‘4‘=1L, ‘5‘=1L)
BookT19 <- timss.mean(variable=’BSBG04’, by=’IDCNTRYL’, data=
   timss19)
colnames(BookT19)[3] <- "BookT19"
BookT19 <- BookT19[c(1,3)]

timss19$BSBG09A <- dplyr::recode(timss19$BSBG09A, ‘1‘=0L,
   ‘2‘=1L)
ImmigT19 <- timss.mean(variable=’BSBG09A’, by=’IDCNTRYL’,
   data=timss19)
colnames(ImmigT19)[3] <- "ImmigT19"
ImmigT19 <- ImmigT19[c(1,3)]

timss19$BCBG13AA <- dplyr::recode(timss19$BCBG13AA, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
ShortT19 <- timss.mean(variable=’BCBG13AA’, by=’IDCNTRYL’,
   data=timss19)
colnames(ShortT19)[3] <- "ShortT19"
ShortT19 <- ShortT19[c(1,3)]

timss19$BCBG13BA <- dplyr::recode(timss19$BCBG13BA, ‘1‘=0L,
   ‘2‘=0L, ‘3‘=1L, ‘4‘=1L)
TeachT19 <- timss.mean(variable=’BCBG13BA’, by=’IDCNTRYL’,
   data=timss19)
colnames(TeachT19)[3] <- "TeachT19"
TeachT19 <- TeachT19[c(1,3)]
```

```
T19 <- list(MeansT19, BookT19, ImmigT19, ShortT19, TeachT19)
T19 <- T19 %>% reduce(full_join, by='IDCNTRYL')
names(T19)[names(T19) == 'IDCNTRYL'] <- 'CNT'
T19$CNT <- dplyr::recode(T19$CNT, "England"="UnitedKingdom")
T19$CNT <- countrycode(T19$CNT, origin = 'country.name',
   destination = 'iso3c')
rm(BookT19, ImmigT19, ShortT19, TeachT19)

rm(MeansT99,MeansT03,MeansT07,MeansT11,MeansT15,MeansT19)
rm(pisa00read,pisa00math,pisa03,pisa06,pisa09,pisa12,pisa15,
   pisa18)
rm(timss99,timss03,timss07,timss11,timss15,timss19)

#Adjustments and selection of countries
MeansT$IDCNTRYL <- as.character(MeansT$IDCNTRYL)
MeansT$IDCNTRYL[MeansT$IDCNTRYL=="England"] <- "United␣
   Kingdom"
MeansT$IDCNTRYL <- countrycode(MeansT$IDCNTRYL, origin = '
   country.name', destination = 'iso3c')
colnames(MeansT)[1] <- "CNT"
MeansT <- MeansT[MeansT$CNT %in% list("AUS","BEL","BGR","GBR"
   ,"ISR","ITA","LVA",
"NLD","NOR","NZL","RUS","SWE","USA"), ]
SdT99 <- SdT99[SdT99$CNT %in% list("AUS","BEL","BGR","GBR","
   ISR","ITA","LVA",
"NLD","NOR","NZL","RUS","SWE","USA"), ]
SdT07 <- SdT07[SdT07$CNT %in% list("AUS","BEL","BGR","GBR","
   ISR","ITA","LVA",
"NLD","NOR","NZL","RUS","SWE","USA"), ]

Sd <- list(SdT99,SdP00,SdP03,SdP06,SdT07,SdP09,SdP12,SdP15)
Sd <- Sd %>% reduce(full_join, by='CNT')
Sd <- Sd[Sd$CNT %in% list("AUS","AUT","BEL","BGR","CAN","CHE"
   ,"CZE",
"DEU","DNK","ESP","FRA","FIN","GBR","IRL",
"ISL","ISR","ITA","LUX","LVA","NLD","NOR",
"NZL","RUS","SWE","USA") ,]
rm(SdT99,SdP00,SdP03,SdP06,SdT07,SdP09,SdP12,SdP15)


# DESCRIPTIVE STATISTICS ----
```

```
#Means statistics (examples)
#StatM99 <- data.frame(MeansT[MeansT$CNT!="ITA",] %>%
#                           dplyr::summarize(donor_sample = mean
   (MeansT99,na.rm = TRUE)),MeansT[MeansT$CNT=="ITA",]$
   MeansT99) #etc.
#StatR00 <- data.frame(Read[Read$CNT!="ITA",] %>%
#                           dplyr::summarize(donor_sample = mean
   (Read00,na.rm = TRUE)),Read[Read$CNT=="ITA",]$Read00) #etc
   .
# P00 <- P00[P15$CNT %in% list("AUS","AUT","BEL","BGR","CAN
   ","CHE","CZE",
#                              "DEU","DNK","ESP","FRA","FIN","GBR
   ","IRL",
#                              "ISL","ISR","ITA","LUX","LVA","NLD
   ","NOR",
#                              "NZL","RUS","SWE","USA") ,]
# StatTeachP00 <- data.frame(P00[P00$CNT!="ITA",] %>%
#                                  dplyr::summarize(donor_sample =
   mean(TeachP00,na.rm = TRUE)),P00[P00$CNT=="ITA",]$TeachP00
   ) #etc.


#SD statistics (examples)
#All zeroes transformed to NaN: some of them are due to real
   NaN, others to 0 means
#(however I pruned both because 0 SD may artificially
   underestimate average SD)
#Sd[Sd == 0] <- NA

#StatSdMT99 <- data.frame(Sd[Sd$CNT!="ITA",] %>%
#                                 dplyr::summarize(donor_sample =
   mean(SdMT99,na.rm = TRUE)),Sd[Sd$CNT=="ITA",]$SdMT99) #etc
   .


# PREPARATION FOR ANALYSIS ----

#I need to impute Bulgaria and Israel's PISA 2003 in order to
    anchor TIMSS 2003
MeansP$MeansP03[which(MeansP$CNT == "BGR")] <-
```

```
   (429.6223+413.4492)/2
MeansP$MeansP03[which(MeansP$CNT == "ISR")] <-
   (432.9734+441.8587)/2

#Merge and select countries
Means <- merge(MeansP,MeansT,by="CNT",all=TRUE)
Means <- Means[, c(1,9,2,3,10,4,11,5,12,6,7,13,8,14)]

#Rescaling (TIMSS 2003 -> PISA 2003)
Means_adj <- Means
Means_adj$coeff <- Means_adj$MeansP03/Means_adj$MeansT03
Means_adj$MeansT99 <- Means_adj$MeansT99*Means_adj$coeff
Means_adj$MeansT03 <- Means_adj$MeansT03*Means_adj$coeff
Means_adj$MeansT07 <- Means_adj$MeansT07*Means_adj$coeff
Means_adj$MeansT11 <- Means_adj$MeansT11*Means_adj$coeff
Means_adj$MeansT15 <- Means_adj$MeansT15*Means_adj$coeff
Means_adj$MeansT19 <- Means_adj$MeansT19*Means_adj$coeff

#Reshaping and pre-imputation adjustments
Means_adj <- as.data.frame(Means_adj)
Means_adj <- reshape(Means_adj,varying=c("MeansT99", "
   MeansP00", "MeansP03",
"MeansT03","MeansP06", "MeansT07", "MeansP09",
"MeansT11", "MeansP12","MeansP15",
"MeansT15", "MeansP18", "MeansT19"),
times=c
   (1999,2000,2003,2003,2006,2007,2009,2011,2012,2015,2015,
2018,2019),
idvar="CNT",drop="coeff",v.names="Score",direction="long")
Means_adj <- Means_adj[order(Means_adj$CNT),]

#The following command simply removes TIMSS03 and TIMSS15
   data
Means_adj <- Means_adj[!duplicated(Means_adj[,c(1,2)]),]
#I truncate at 2015 (notice this is done before imputation)
Means_adj <- Means_adj[Means_adj$time != 2018 & Means_adj$
   time != 2019, ]
#Remove PISA 2000 before imputation (no comparability)
PisaMath2000 <- Means_adj[Means_adj$time==2000, ]
Means_adj <- Means_adj[Means_adj$time!=2000, ]
```

```
#Reading
Read <- as.data.frame(Read)
Read <- reshape(Read,varying=c("Read00", "Read03","Read06","
   Read09",
"Read12","Read15", "Read18"),
times=c(2000,2003,2006,2009,2012,2015,2018),
idvar="CNT",v.names="Score",direction="long")
Read <- Read[order(Read$CNT),]
#Truncate reading scores at 2015
Read <- Read[Read$time != 2018, ]

#Imputation (w/ linear interpolation)
Means_adj_1 <- Means_adj
for (i in 1:25) {
        Means_adj_1[c((1+8*(i-1)):(4+8*(i-1))),3] <-
        na_interpolation(Means_adj_1[c((1+8*(i-1)):(4+8*(i-1)
           )),3], option = "linear")}
for (i in 1:25) {
        Means_adj_1[c((5+8*(i-1)):(8+8*(i-1))),3] <-
        na_interpolation(Means_adj_1[c((5+8*(i-1)):(8+8*(i-1)
           )),3], option = "linear")}

#Alternative imputation (w/ linear regression)
Means_adj_2 <- Means_adj
for (i in 1:25) {
        Means_adj_2[c((1+8*(i-1)):(4+8*(i-1))),c(3,2)] <-
        regressionImp(Score~time,data=Means_adj_2[c((1+8*(i
           -1)):(4+8*(i-1))),c(3,2)],imp_var="FALSE")}
for (i in 1:25) {
        Means_adj_2[c((5+8*(i-1)):(8+8*(i-1))),c(3,2)] <-
        regressionImp(Score~time,data=Means_adj_2[c((5+8*(i
           -1)):(8+8*(i-1))),c(3,2)],imp_var="FALSE")}
rm(i)

#Pisa-only reading imputation (w/ linear regression)
for (i in 1:25) {
        Read[c((1+6*(i-1)):(3+6*(i-1))),c(2,3)] <-
           regressionImp(Score~time,data=Read[c((1+6*(i-1))
           :(3+6*(i-1))),c(2,3)],imp_var="FALSE")}
for (i in 1:25) {
        Read[c((4+6*(i-1)):(6+6*(i-1))),c(2,3)] <-
```

```
          regressionImp(Score~time,data=Read[c((4+6*(i-1))
          :(6+6*(i-1))),c(2,3)],imp_var="FALSE")}


#Predictors for PISA-TIMSS merged dataset
PTp <- list(P00,P03,P06,P09,P12,P15,P18)
PTp <- PTp %>% reduce(full_join, by='CNT')
PTt <- list(T99,T03,T07,T11,T15,T19)
PTt <- PTt %>% reduce(full_join, by='CNT')
PTt <- PTt[PTt$CNT %in% list("AUS","BEL","BGR","GBR","ISR","
   ITA","LVA",
"NLD","NOR","NZL","RUS","SWE","USA") ,]
PT <- merge(PTp,PTt,by="CNT",all=TRUE)
rm(PTp,PTt)
PT <- PT %>% dplyr::select(-(contains("Means") | contains("
   Private") | contains("Account") | contains("Stratio")))
PT <- PT[, c
   (1,30,31,32,33,2,3,4,5,6,7,8,9,34,35,36,37,10,11,12,13,38,
39,40,41,14,15,16,17,42,43,44,45,18,19,20,21,22,23,24,25,46,47,

48,49,26,27,28,29,50,51,52,53)]
PT <- PT[PT$CNT %in% list("AUS","AUT","BEL","BGR","CAN","CHE"
   ,"CZE",
"DEU","DNK","ESP","FRA","FIN","GBR","IRL",
"ISL","ISR","ITA","LUX","LVA","NLD","NOR",
"NZL","RUS","SWE","USA") ,]
PT <- reshape(PT, idvar="CNT", direction="long",
varying=list(Book=c(2,6,10,14,18,22,26,30,34,38,42,46,50),
Immig=c(3,7,11,15,19,23,27,31,35,39,43,47,51),
Short=c(4,8,12,16,20,24,28,32,36,40,44,48,52),
Teach=c(5,9,13,17,21,25,29,33,37,41,45,49,53)),
times=c
   (1999,2000,2003,2003,2006,2007,2009,2011,2012,2015,2015,
2018,2019),
v.names = c("Book", "Immig","Short","Teach"))
PT <- PT[order(PT$CNT),]
PT <- PT[!duplicated(PT[,c(1,2)]),]
#Here, I truncate at 2015.
PT <- PT[PT$time != 2018 & PT$time != 2019, ]
#Here, I interpolate over all years because my assumption is
   that predictors are not affected by the policy.
```

```
#Notice that only values before 2009 are used to compute SCM
   weights.
PT$Book <- as.numeric(PT$Book)
for (i in 1:25) {
        PT$Book[((1+9*(i-1)):(9+9*(i-1)))] <- na_
           interpolation(PT$Book[((1+9*(i-1)):(9+9*(i-1)))],
           option = "linear")}
PT$Immig <- as.numeric(PT$Immig)
for (i in 1:25) {
        PT$Immig[((1+9*(i-1)):(9+9*(i-1)))] <- na_
           interpolation(PT$Immig[((1+9*(i-1)):(9+9*(i-1)))],
            option = "linear")}
PT$Short <- as.numeric(PT$Short)
for (i in 1:25) {
        PT[c((1+9*(i-1)):(5+9*(i-1))),c(5,2)] <-
        regressionImp(Short~time,data=PT[c((1+9*(i-1)):(5+9*(
           i-1))),c(5,2)],imp_var="FALSE")}
for (i in 1:25) {
        PT[c((6+9*(i-1)):(9+9*(i-1))),c(5,2)] <-
        regressionImp(Short~time,data=PT[c((6+9*(i-1)):(9+9*(
           i-1))),c(5,2)],imp_var="FALSE")}
PT$Teach <- as.numeric(PT$Teach)
for (i in 1:25) {
        PT[c((1+9*(i-1)):(5+9*(i-1))),c(6,2)] <-
        regressionImp(Teach~time,data=PT[c((1+9*(i-1)):(5+9*(
           i-1))),c(6,2)],imp_var="FALSE")}
for (i in 1:25) {
        PT[c((6+9*(i-1)):(9+9*(i-1))),c(6,2)] <-
        regressionImp(Teach~time,data=PT[c((6+9*(i-1)):(9+9*(
           i-1))),c(6,2)],imp_var="FALSE")}

PT <- merge(PT,gdp,by=c("CNT","time"),all.x=TRUE)
PT <- merge(PT,private,by=c("CNT","time"),all.x=TRUE)

#Predictors for PISA-only dataset
P <- list(P00,P03,P06,P09,P12,P15,P18)
P <- P %>% reduce(full_join, by=’CNT’)
P <- P %>% dplyr::select(-(contains("Means") | contains("
   Private") | contains("Teach")))
P <- P[P$CNT %in% list("AUS","AUT","BEL","BGR","CAN","CHE","
   CZE",
"DEU","DNK","ESP","FRA","FIN","GBR","IRL",
"ISL","ISR","ITA","LUX","LVA","NLD","NOR",
"NZL","RUS","SWE","USA") ,]
P <- reshape(P, idvar="CNT", direction="long",
varying=list(Book=c(2,7,12,17,22,27,32),
Immig=c(3,8,13,18,23,28,33),
Short=c(4,9,14,19,24,29,34),
Account=c(5,10,15,20,25,30,35),
Stratio=c(6,11,16,21,26,31,36)),
times=c(2000,2003,2006,2009,2012,2015,2018),
v.names = c("Book","Immig","Short","Account","Stratio"))
P <- P[order(P$CNT),]
#Here, I truncate at 2015.
P <- P[P$time != 2018, ]
#Here, I interpolate over all years because my assumption is
   that predictors are not affected by the policy.
#Notice that only values before 2009 are used to compute SCM
   weights.
P$Book <- as.numeric(P$Book)
for (i in 1:25) {
        P$Book[((1+6*(i-1)):(6+6*(i-1)))] <- na_interpolation
           (P$Book[((1+6*(i-1)):(6+6*(i-1)))], option = "
           linear")
}
P$Immig <- as.numeric(P$Immig)
for (i in 1:25) {
        P$Immig[((1+6*(i-1)):(6+6*(i-1)))] <- na_
           interpolation(P$Immig[((1+6*(i-1)):(6+6*(i-1)))],
           option = "linear")
}
P$Account <- as.numeric(P$Account)
for (i in 1:25) {
        P$Account[((1+6*(i-1)):(6+6*(i-1)))] <- na_
           interpolation(P$Account[((1+6*(i-1)):(6+6*(i-1)))
           ], option = "linear")
}
P$Short[is.nan(P$Short)]<-NA
P$Short <- as.numeric(P$Short)
for (i in 1:25) {
        P[c((1+6*(i-1)):(3+6*(i-1))),c(5,2)] <-
        regressionImp(Short~time,data=P[c((1+6*(i-1)):(3+6*(i
            -1))),c(5,2)],imp_var="FALSE")}
for (i in 1:25) {
        P[c((4+6*(i-1)):(6+6*(i-1))),c(5,2)] <-
        regressionImp(Short~time,data=P[c((4+6*(i-1)):(6+6*(i
            -1))),c(5,2)],imp_var="FALSE")}
P$Stratio[is.nan(P$Stratio)]<-NA
P$Stratio <- as.numeric(P$Stratio)
for (i in 1:25) {
        P[c((1+6*(i-1)):(3+6*(i-1))),c(7,2)] <-
        regressionImp(Stratio~time,data=P[c((1+6*(i-1)):(3+6*
            (i-1))),c(7,2)],imp_var="FALSE")}
for (i in 1:25) {
        P[c((4+6*(i-1)):(6+6*(i-1))),c(7,2)] <-
        regressionImp(Stratio~time,data=P[c((4+6*(i-1)):(6+6*
            (i-1))),c(7,2)],imp_var="FALSE")}
gdpRead <- gdp[gdp$time %in% c(2000,2003,2006,2009,2012,2015)
   ,]
privateRead <- private[private$time %in% c
   (2000,2003,2006,2009,2012,2015),]
P <- merge(P,gdpRead,by=c("CNT","time"),all.x=TRUE)
P <- merge(P,privateRead,by=c("CNT","time"),all.x=TRUE)
rm(gdp,private,gdpRead,privateRead)

#Final datasets
SCMpt <- merge(Means_adj_2,PT,by=c("CNT","time"),all.x=TRUE)
SCMp <- merge(Read,P,by=c("CNT","time"),all.x=TRUE)
pm <- Means_adj_2[Means_adj_2$time != 1999 & Means_adj_2$time
    != 2007 & Means_adj_2$time != 2011, ]
SCMpM <- merge(pm,P,by=c("CNT","time"),all.x=TRUE)
rm(pm)
#Remove TIMSS 2011
SCMpt <- SCMpt[SCMpt$time != 2011, ]

#Donor pool approximation
ApproxM <- data.frame(SCMpM[SCMpM$CNT!="ITA",] %>% dplyr::
   group_by(time) %>% dplyr::summarize(donor_sample = mean(
   Score)),SCMpM[SCMpM$CNT=="ITA",]$Score)
colnames(ApproxM) <- c("Time","Donor","Italy")
ApproxM %>% plot_approxM()

ApproxR <- data.frame(SCMp[SCMp$CNT!="ITA",] %>% dplyr::group
   _by(time) %>% dplyr::summarize(donor_sample = mean(Score))
   ,SCMp[SCMp$CNT=="ITA",]$Score)
colnames(ApproxR) <- c("Time","Donor","Italy")
ApproxR %>% plot_approxR()

# DATA ANALYSIS ----

#PISA Math only
SCMpM_out <-
SCMpM %>% synthetic_control(outcome = Score, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 2003:2006,
book = mean(Book, na.rm = TRUE),
immig = mean(Immig, na.rm = TRUE),
account = mean(Account, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
) %>%
generate_predictor(time_window = 2003, score_2003 = Score)
   %>%
generate_predictor(time_window = 2006, score_2006 = Score)
   %>%
generate_weights(optimization_window = 2003:2006#, margin_
   ipop = .02,sigf_ipop = 9,bound_ipop = 6
) %>%
generate_control()
SCMpM_out %>% plot_trends2()
SCMpM_out %>% plot_weights()
SCMpM_out %>% grab_unit_weights() %>% print(n = Inf)
SCMpM_out %>% grab_predictor_weights() %>% print(n = Inf)
SCMpM_out %>% grab_balance_table()
SCMpM_out %>% plot_placebos2(prune = FALSE)
SCMpM_out %>% plot_placebos2(prune = TRUE)
SCMpM_out %>% plot_mspe_ratio()
SCMpM_out %>% grab_significance() %>% print(n = Inf)
```

```
#PISA Read only
SCMp_out <-
SCMp %>% synthetic_control(outcome = Score, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 2000:2006,
book = mean(Book, na.rm = TRUE),
immig = mean(Immig, na.rm = TRUE),
account = mean(Account, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
) %>%
generate_predictor(time_window = 2000, score_2000 = Score)
   %>%
generate_predictor(time_window = 2003, score_2003 = Score)
   %>%
generate_predictor(time_window = 2006, score_2006 = Score)
   %>%
generate_weights(optimization_window = 2000:2006#, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()
SCMp_out %>% plot_trends3()
SCMp_out %>% plot_weights()
SCMp_out %>% grab_unit_weights() %>% print(n = Inf)
SCMp_out %>% grab_predictor_weights() %>% print(n = Inf)
SCMp_out %>% grab_balance_table()
SCMp_out %>% plot_placebos3(prune = FALSE)
SCMp_out %>% plot_placebos3(prune = TRUE)
SCMp_out %>% plot_mspe_ratio()
SCMp_out %>% grab_significance() %>% print(n = Inf)

#PISA and TIMSS (Math)
SCMpt_out <-
SCMpt %>% synthetic_control(outcome = Score, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 1999:2007,
book = mean(Book, na.rm = TRUE),
immig = mean(Immig, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
)%>%
generate_predictor(time_window = 1999, score_1999 = Score)
   %>%
generate_predictor(time_window = 2003, score_2003 = Score)
   %>%
generate_predictor(time_window = 2006, score_2006 = Score)
   %>%
generate_predictor(time_window = 2007, score_2007 = Score)
   %>%
generate_weights(optimization_window = 1999:2007#, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()
SCMpt_out %>% plot_trends4()
SCMpt_out %>% plot_differences()
SCMpt_out %>% plot_weights()
SCMpt_out %>% grab_unit_weights() %>% print(n = Inf)
SCMpt_out %>% grab_predictor_weights() %>% print(n = Inf)
SCMpt_out %>% grab_balance_table()
SCMpt_out %>% plot_placebos4(prune = FALSE)
SCMpt_out %>% plot_placebos4(prune = TRUE)
SCMpt_out %>% plot_mspe_ratio()
SCMpt_out %>% grab_significance() %>% print(n = Inf)


# ROBUSTNESS ----

#PISA Math only and until 2003
SCMpM_out_03 <-
```

```
SCMpM %>% synthetic_control(outcome = Score, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 2003,
book = mean(Book, na.rm = TRUE),
immig = mean(Immig, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
)%>%
generate_predictor(time_window = 2003, score_2003 = Score)
   %>%
generate_weights(optimization_window = 2003#, margin_ipop =
   .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()
SCMpM_out_03 %>% plot_trends2()
SCMpM_out_03 %>% plot_differences()
SCMpM_out_03 %>% plot_weights()
SCMpM_out_03 %>% grab_balance_table()
SCMpM_out_03 %>% plot_placebos_validation(prune = FALSE)
SCMpM_out_03 %>% plot_placebos_validation(prune = TRUE)
SCMpM_out_03 %>% plot_mspe_ratio()
SCMpM_out_03 %>% grab_significance() %>% print(n = Inf)

#Until 2003 Read
SCMp_out_03 <-
SCMp %>% synthetic_control(outcome = Score, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 2000:2003,
book = mean(Book, na.rm = TRUE),
immig = mean(Immig, na.rm = TRUE),
account = mean(Account, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
) %>%
generate_predictor(time_window = 2000, score_2000 = Score)
   %>%
generate_predictor(time_window = 2003, score_2003 = Score)
   %>%
generate_weights(optimization_window = 2000:2003 #, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()
SCMp_out_03 %>% plot_trends()
SCMp_out_03 %>% plot_weights()
SCMp_out_03 %>% plot_placebos(prune = FALSE)
SCMp_out_03 %>% plot_placebos(prune = TRUE)
SCMp_out_03 %>% plot_mspe_ratio()
SCMp_out_03 %>% grab_significance() %>% print(n = Inf)

#Match until 2003 only (PISA + TIMSS)
SCMpt_out_03 <-
SCMpt %>% synthetic_control(outcome = Score, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 1999:2003,
book = mean(Book, na.rm = TRUE),
immig = mean(Immig, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
)%>%
generate_predictor(time_window = 1999, score_1999 = Score)
   %>%
generate_predictor(time_window = 2003, score_2003 = Score)
   %>%
generate_weights(optimization_window = 1999:2003#, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()
SCMpt_out_03 %>% plot_trends()
SCMpt_out_03 %>% plot_differences()
SCMpt_out_03 %>% plot_weights()
SCMpt_out_03 %>% grab_balance_table()
SCMpt_out_03 %>% plot_placebos(prune = FALSE)
SCMpt_out_03 %>% plot_placebos(prune = TRUE)
SCMpt_out_03 %>% plot_mspe_ratio()
SCMpt_out_03 %>% grab_significance() %>% print(n = Inf)

#Eliminate one country at a time (USA, ISR, BGR)
#--> substantially similar to baseline (without ISR the fit
   is worse; without BGR the jump is in 2007)
SCMpMm <- SCMpM[SCMpM$CNT != "ISR", ]
SCMpMm_out <-
SCMpMm %>% synthetic_control(outcome = Score, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 2003:2006,
book = mean(Book, na.rm = TRUE),
immig = mean(Immig, na.rm = TRUE),
account = mean(Account, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
)%>%
generate_predictor(time_window = 2003, score_2003 = Score)
   %>%
generate_predictor(time_window = 2006, score_2006 = Score)
   %>%
generate_weights(optimization_window = 2003:2006#, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()
SCMpMm_out %>% plot_trends2()
SCMpMm_out %>% plot_weights()
SCMpMm_out %>% plot_placebos2(prune = FALSE)
SCMpMm_out %>% plot_mspe_ratio()

#Re-run with Means_adj_1
#--> very similar results

#Re-run without one predictor at a time
#--> same results



# MECHANISMS ----
Stratio_out <-
SCMp %>% synthetic_control(outcome = Stratio, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 2000:2006,
score = mean(Score, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
) %>%
generate_predictor(time_window = 2000, stratio_2000 = Stratio
   ) %>%
generate_predictor(time_window = 2003, stratio_2003 = Stratio
   ) %>%
generate_predictor(time_window = 2006, stratio_2006 = Stratio
   ) %>%
generate_weights(optimization_window = 2000:2006#, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()

Stratio_out %>% plot_trends()
Stratio_out %>% plot_weights()
Stratio_out %>% plot_placebos_stratio(prune = FALSE)
Stratio_out %>% plot_placebos_stratio(prune = TRUE)
Stratio_out %>% plot_mspe_ratio()


Teach_out <-
SCMpt %>% synthetic_control(outcome = Teach, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
i_time = 2008, # time period when the intervention occurred
generate_placebos=TRUE # generate placebo synthetic controls
   (for inference)
) %>%
generate_predictor(time_window = 1999:2007,
score = mean(Score, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
)%>%
generate_predictor(time_window = 1999, teach_1999 = Teach)
   %>%
generate_predictor(time_window = 2003, teach_2003 = Teach)
   %>%
generate_predictor(time_window = 2006, teach_2006 = Teach)
   %>%
generate_predictor(time_window = 2007, teach_2007 = Teach)
   %>%
generate_weights(optimization_window = 1999:2007#, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()

Teach_out %>% plot_trends()
Teach_out %>% plot_weights()
Teach_out %>% plot_placebos_teach(prune = FALSE)
Teach_out %>% plot_placebos_teach(prune = TRUE)
Teach_out %>% plot_mspe_ratio()

Short_out <-
SCMp %>% synthetic_control(outcome = Short, # outcome
unit = CNT, # unit index in the panel data
time = time, # time index in the panel data
i_unit = "ITA", # unit where the intervention occurred
```

```
i_time = 2008, # time period when the intervention occurred
generate_placebos = TRUE # generate placebo synthetic
   controls (for inference)
) %>%
generate_predictor(time_window = 2000:2006,
score = mean(Score, na.rm = TRUE),
gdp = mean(GDPpc, na.rm = TRUE),
private = mean(Private, na.rm = TRUE)
)%>%
generate_predictor(time_window = 2000, short_2000 = Short)
   %>%
generate_predictor(time_window = 2003, short_2003 = Short)
   %>%
generate_predictor(time_window = 2006, short_2006 = Short)
   %>%
generate_weights(optimization_window = 2000:2006#, margin_
   ipop = .02,sigf_ipop = 7,bound_ipop = 6
) %>%
generate_control()
Short_out %>% plot_trends()
Short_out %>% plot_weights()
Short_out %>% plot_placebos_short(prune = FALSE)
Short_out %>% plot_placebos_short(prune = TRUE)
Short_out %>% plot_mspe_ratio()


# PENALIZED SC ----
#Functions for penalized SC available from https://github.com
   /jeremylhour/pensynth
setwd("~/Desktop/UNIVERSITY/Economics/2 ␣Anno/Master's␣
   Thesis/Data␣analysis/Functions")
source("wsoll1.R")
source("regsynth.R")
source("TZero.R")
source("estimator_matching.R")
source("get_stats.R")

SCMpM$n <- rep(1:25,each=5)
time <- c(2003,2006,2009,2012,2015)
final <- data.frame(time)
for (i in (1:25)) {
        prova.out <- dataprep(foo = SCMpM,
        predictors = c("Book", "Immig", "Account"),
        predictors.op = "mean",
        time.predictors.prior = c(2003,2006),
        special.predictors = list(list("Score",2003,"mean"),
        list("Score",2006, "mean")),
        dependent = "Score",
        unit.variable = "n",
        unit.names.variable = "CNT",
        time.variable = "time",
        treatment.identifier = i,
        controls.identifier = c(1:25)[-i],
        time.optimize.ssr = c(2003,2006),
        time.plot = c(2003,2006,2009,2012,2015))
        sol <- regsynth(prova.out$X0,prova.out$X1,colMeans(
           prova.out$Y0plot),colMeans(prova.out$Y1plot),pen
           =.1,parallel=TRUE)
        col_name <- paste0('A',i)
        df_temp <- data.frame(add=SCMpM[SCMpM$n==i,3] - round
           (apply(prova.out$Y0plot%*%t(sol$Wsol),1,mean),
           digits=2))
        colnames(df_temp) <- col_name
        final <- cbind(final,df_temp)
}
final <- melt(final ,  id.vars = 'time', variable.name = '
   Differences')
final$dummy <- ifelse(final$Differences=="A17",1,0)
final %>% plot_penalized()

final$sqr <- (final$value)^2
final <- final %>% dplyr::group_by(Differences) %>%
mutate(preRMSPE = (((sqr[1]+sqr[2])/2)^(1/2)))
final <- final %>% dplyr::group_by(Differences) %>%
mutate(postRMSPE = (((sqr[3]+sqr[4]+sqr[5])/3)^(1/2)))
final$RMSE <- final$postRMSPE/final$preRMSPE
```