A New Horizon? The Effect of a National Education Reform on Student Achievement and the Academic Environment

Yaniv Reingewertz and Adi Shany*

March 2024

Abstract

This study analyzes the effect of a whole-school national educational reform in Israel on student achievements and the academic environment in schools. The New Horizon reform was initiated by the Ministry of Education and gradually implemented in elementary schools beginning in 2008. The reform increased the time teachers spend in school, added more small-group instruction time (individual hours), increased teachers' and principals' salaries, and established higher requirements and incentives for professional development. We use a staggered difference-in-differences approach to evaluate the effects of the reform on student achievement and the school learning environment. We show that over the three years following the introduction of the reform, test scores rose in math and English, and improvements were seen in various aspects of the school learning environment. We offer suggestive evidence indicating that the short-term effects of the reform are likely driven by small-group learning, improved working conditions for certain subject teachers, and an increase in the number of hours teachers spend in school.

Keywords: School reform, student achievements, school learning environment

JEL Codes: I21, I24, I28

*Yaniv Reingewertz is an Associate Professor at the Division of Public Administration and Policy, School of Political Sciences, The University of Haifa. Adi Shany is an Assistant Professor at the Coller School of Management, Tel Aviv University. We thank the Israel's Ministry of Education and the National Authority for Measurement and Evaluation in Education (RAMA) for allowing access to restricted data at the Ministry online protected research lab. Many thanks to Haim Gat, Sonia Perez, Orna Simhon, and Mira Lahak who provided much insight and information on the New Horizon reform. We thank Eliad Trefler and Ariela Knaani from the Ministry online protected research lab for their great assistance. Shany acknowledges financial support by the Henry Crown Institute of Business Research in Israel. Contact email: adishany@tauex.tau.ac.il .

1 Introduction

Improving student outcomes through school reforms is a central focus in education economics policy. Key questions revolve around the conditions under which these reforms succeed, how their success is measured, and the necessary inputs for their effectiveness. Despite their significance for education policy decisions, there has been a lack of comprehensive evaluation of nationwide education reforms.[1] Existing literature primarily focuses on how specific inputs (such as class size, instructional time, teacher quality) impact student achievement.[2] This study seeks to address this gap by analyzing the impact of a nationwide education reform in Israel on student achievement and the school learning environment, as well as providing insights into the potential mechanisms driving the reform effects.

We study the effect of a whole-school broad-based national reform introduced in Israeli elementary schools in 2008. This comprehensive intervention which still operates today, known as the New Horizon (NH) reform, is the largest investment in Israel's education system within the last few decades. The reform's main objectives were to improve teachers' working conditions, improve the learning environment, reduce educational gaps, and raise overall student achievement (RAMA, 2012a, RAMA, 2012b). The core of the reform measures included introducing small-group learning, increasing teachers' working hours at school alongside raising teachers' salaries, and introducing new criteria governing teachers' professional development and promotion in order to make the profession more attractive to talented newcomers and encourage greater engagement among existing teachers.

In this paper, we present an empirical investigation of how the NH reform affected

[1] An exception worth noting is Dee and Jacob (2011) who examines the impact of the No Child Left Behind law in the US. Additionally, there is an extensive literature on the effects of school finance reforms (for example Jackson et al., 2016, Hyman, 2017, Lafortune et al., 2018a, Lafortune et al., 2018b, Johnson and Jackson, 2019, Jackson, 2020, Jackson, 2021, Jackson and Mackevicius, 2021, Baron, 2022, Biasi, 2023 and Biasi et al., 2024).

[2] See among others Angrist and Lavy (1999), Hoxby (2000), Jepsen and Rivkin (2009), Chetty et al. (2011), Harris and Sass (2011), Rivkin and Schiman (2015), Araujo et al. (2016), Justman (2018), Angrist et al. (2019), Leuven and Løkken (2020), Gilraine (2020), Lavy (2020), Kedagni et al. (2021), Biasi (2021), Barrios-Fernández and Bovini (2021).

educational and school learning environment outcomes in the short term, using a sample which covers all Jewish state elementary schools in Israel over the years 2005–2012.[3] Our outcome data are drawn from fifth-grade test scores on national standardized exams in math, language (i.e. reading and writing in Hebrew), science, and English, along with questionnaires polling fifth- and sixth-graders on their school learning environment, and questionnaires distributed to all teachers and principals in participating schools. We merge this data with demographic information on students, teachers, and schools. This rich dataset allows us to explore multiple dimensions of the reform and its effect on students, teachers, and principals.

Our identification strategy relies on the fact that the assignment of schools into the NH reform program was gradual. Given this staggered implementation, we use a staggered difference-in-differences estimation approach to compare educational and school learning environment outcomes before and after the reform. The implementation of the NH reform was not random, and some schools made an active choice to join the reform early. However, the key identifying assumption underlying this research design is that the exact timing of the reform is unrelated to potential outcomes. This assumption implies that the outcomes for early- and late-reforming schools would have trended similarly in the absence of the reform. Our event study analysis shows that, indeed, key outcomes trended similarly in the years before the reform. We also provide evidence that our results are not driven by a cohort of schools that implemented the reforms in a given year and that pre-determined students' and schools' characteristics do not correlate with the timing of the reform. To address recent concerns raised by the econometric literature on DiD models with staggered adoption, we show that our results are robust to using an alternative estimator proposed by Sun and Abraham (2021).

The results suggest that over the study period, the NH reform has had a positive and significant short-run effect on both math and English test scores and the school learning environment. However, test scores in language and science did not improve as a result

[3] Arab schools and Ultra-Orthodox schools are operating as separate streams and are thus not included in our sample.

of the NH reform. In Math and English, the effect is not immediate, as it only begins in the second year of the reform and ranges from 0.1 to 0.2 standard deviations increase. In terms of environmental outcomes, the reform improved the relationship between teachers and students, increased teachers' efforts to help their students, and improved students' behavior in class.

Probing more deeply, we find a larger effect in math test scores among students whose parents had education levels below the median, used as a proxy for students' ability. The latter group also showed improvement in language test scores. These findings may be attributed to the greater likelihood of lower-achieving students being included in small-group learning sessions, which primarily focus on math and language subjects. However, the effect of the reform on English test scores does not vary between students with different levels of parents' education. We explore additional potential explanations for our results by examining teacher data. First, we analyze whether the reform changed the composition of teachers within schools in terms of tenure and holding an academic degree. Then we evaluate differences in working hours and workload by teachers' specialization, and provide suggestive evidence that the effect on English test scores may reflect greater improvement in the working conditions of English teachers after the reform.

The NH reform encompasses a bundle of interventions commonly found in education systems worldwide, such as small-group learning and teacher evaluation in "No Excuses" charter schools and policy interventions that led to changes in teacher salary, whether conditional on teacher attributes or not (Hendricks, 2015, De Ree et al., 2018 and Biasi, 2021). Our analysis provides indicative insights into the combined impact of similar interventions within the context of a nationwide educational reform. The pattern of our results is consistent with previous studies on whole-school policy interventions in the US. Studies on charter schools and the Every Student Succeeds Act reveal that these policies increase student achievement mainly in math and that interventions that include small-group instruction and frequent feedback for teachers produced larger positive effects (Fryer, 2014,

Chabrier et al., 2016, Cohodes and Parham, 2021, Schueler et al., 2022).[4]

Our paper primarily contributes to the literature on school reforms. National reform initiatives are a central focus of interest in both policy and practice. However, these reforms remain understudied, primarily due to the challenge of identifying their causal effects. The existing literature mainly focuses on the effects of school finance reforms which aim to improve outcomes by increasing overall school spending.[5] Yet, the question of what kinds of spending increases matter the most remains unresolved.[6] Another strand of the literature evaluates the effect of different whole-school policy interventions, such as turnaround or charter conversion, that target specific districts or schools (see review by Schueler et al., 2022). Our study provides the opportunity to assess the causal impact of a comprehensive nationwide "whole-school" reform that extends to nearly the entire elementary school system. Therefore, the findings may have broader general implications. Moreover, the richness of our data enables us to provide suggestive evidence regarding the mechanisms by which the reform affects student outcomes. Our suggestive findings add to the review conducted by Schueler et al. (2022), who shed light on the components of successful whole-school policy interventions aimed at improving low-performing schools. Our study explores the effectiveness of nationwide intervention components and provides insights into the role of small-group learning and teacher working conditions in generating academic gains.

Lastly, and importantly, our paper evaluates the effect of the reform on nonacademic outcomes such as the school learning environment—an aspect often overlooked in many studies of educational interventions that focus solely on academic outcomes. Nonacademic outcomes are important, as they are closely tied to non-test outcomes which predict longer-term outcomes such as high-school completion and starting college (Jackson,

[4]Note that we find only moderate effect on language test score for students with low educated parents which is consistent with findings on English Language Arts (ELA) in many studies on the US (Schueler et al., 2022).

[5]For example Jackson et al. (2016), Hyman (2017), Lafortune et al. (2018a), Lafortune et al. (2018b), Johnson and Jackson (2019), Jackson (2020), Jackson (2021), Jackson and Mackevicius (2021), Baron (2022), and Biasi (2023)).

[6]Recent work by Biasi et al. (2024) takes a step toward filling this gap by identifying which investments in school facilities help students.

2018). To the best of our knowledge, this is one of the first attempts to evaluate the effect of school reforms on non-academic outcomes related to student–teacher relationships and violence in school (operationalized as physical bullying and social bullying).[7]

The rest of this paper is organized as follows. The next section provides institutional background on the reform. Section 3 describes the data and Section 4 the empirical strategy. Section 5 presents the results. Section 6 discusses potential channels for the effect of the reform, and Section 7 concludes.

2 Institutional Background

The NH educational reform agreement was signed in March 2008 between the Israeli Ministry of Education and the teacher's union, which represented all state elementary and some lower-secondary school teachers.[8] This was a significant reform which still operates within the Israeli education system today. It increased funding to the education system by 5% yearly, which is the sharpest rise in education funding within the last few decades. All target schools were fully funded by the state, encompassing Hebrew-speaking secular state schools, state religious schools, and Arabic-speaking state schools, which are 75% of all elementary schools in Israel.

The reform was implemented gradually, with roughly 20% of the elementary schools enrolling each year from 2008 to 2012. The reform objectives were to improve teacher's working conditions, improve the learning environment, reinforce and broaden the scope of the school principal's authority, reduce educational gaps, and raise overall student achievement (RAMA, 2012a, RAMA, 2012b). Table 1 describes the number of elementary schools participating in the reform each year by stream. In the 2008/2009 school year, the reform was implemented in 301 schools, all of which had voluntarily adopted the reform before the final agreement was signed.[9] Beginning with the reform's second

[7] Gleason et al. (2010), in their study of charter schools, also evaluate non-academic outcomes such as student effort, behavior, and attitudes in school, as well as parental involvement and satisfaction. According to the review by**?**, only a small number of studies have analyzed non-test outcomes such as attendance, discipline, and graduation.

[8] The reform was not implemented in private and ultra-Orthodox religious schools.

[9] Forty percent of these schools had already undergone a previous school reform, and many strongly

Table 1: Number of Schools Joining the New Horizon Reform by Year and Stream

School year	Hebrew-speaking state schools (Jewish secular)	Hebrew-speaking state schools (Jewish religious)	Arabic-speaking state schools	Total
	(1)	(2)	(3)	(4)
2007/2008	144	44	113	301
2008/2009	241	87	95	423
2009/2010	224	105	74	403
2010/2011	142	94	105	341
2011/2012	45	49	30	124
2012/2013	7	2	3	12

Notes: This table reports the number of schools that joined the New Horizon reform in each school year by stream.

year (the 2009/2010 school year), schools were selected to participate in the reform by school district managers, based on district quotas established by the Ministry of Education.[10] At this stage, once a school joined the reform, all teachers in the school were bound by the new terms immediately upon the reform's implementation. By the beginning of the 2011/2012 school year, 1,468 elementary schools had joined the reform, and since the 2012/2013 school year almost all of Israel's elementary schools have been operating under the post-reform rules.

The main reform measures included introducing small-group learning, increasing teachers' working hours at school alongside raising teachers' salaries, and introducing new criteria governing teachers' professional development and promotion. The implementation of small-group learning, an educational approach wherein approximately 3-5 students form a study group, was an important facet of the reform. Conducted during regular school hours, this method entails withdrawing students from their main classes for fo-

supported the educational vision embodied by the NH reform (RAMA, 2008).

[10]The Israeli education system is divided into seven districts, with district managers responsible for implementing the reform in their district. In July 2016 we interviewed the head of the northern education district, who had held the post during the first years of the NH implementation. According to her, the schools were selected based on the willingness of the teachers and the principal to adopt the reform. In rare cases involving teachers whose schools adopted the reform but who did not themselves wish to join, special arrangements were found (such as an early retirement, etc.). These teachers could not usually transfer to other schools because the reform had become mandatory countrywide by 2012.

cused tutoring sessions. Primarily designed to enhance the academic performance of specific students, often those struggling in subjects like math and Hebrew language, small-group learning, as observed in practice, displayed a broader diversity, including students from various achievement levels and covering additional subjects, albeit to a lesser extent (RAMA, 2008). This approach not only fostered a more personalized learning environment but also significantly improved the social and emotional connections between students and teachers (RAMA, 2008).

The reform included an increase in the work week of full-time teachers from 26 to 36 hours. As was the case before the reform, 26 hours were to be devoted to regular instruction. The ten additional hours were divided between small-group instruction and teaching-related activities (e.g., activities such as class preparation or marking that teachers had previously performed at home).[11] In return for the increased workload, teachers received a wage increase averaging twenty-six percent (Cohen, 2011).[12] Additional pay grades were added to allow teachers more opportunities for promotion in return for professional development, and a bachelor's degree and a teaching certificate were made conditions of employment under the reform (previously only a teaching certificate was needed). The reform was intended to address what many perceived as a crisis in teacher morale and the challenge of attracting high-quality entrants into the profession.[13] The pay and responsibilities of principals also underwent changes, with a separate and more generous pay scale introduced. Principals were granted greater control over hiring, tenure, promotion, and the ability to initiate procedures for firing teachers. Additionally, a special training college for principals was established. The reform also mandated improvements in teachers' physical environment, in particular the provision of suitable work areas.

[11] The figures of 26 hours and 10 additional hours relate only to full-time teachers. For part-time teachers (a large proportion of Israel's teaching population), these figures are reduced proportionately.

[12] These salary hikes were particularly substantial for junior staff, with the starting salary for new teachers nearly doubling and pay for veteran teachers increasing by about one quarter (Hemmings, 2010).

[13] Before the reform, in 2007, the average pay of teachers in Israel after 15 years of experience was equivalent to 62 percent of GDP per capita. In 2013, five years after the reform's initial implementation, the average was equivalent to almost 100 percent of GDP per capita. According to OECD (2009) and OECD (2015), 100 percent of GDP per capita falls toward the bottom end of the typical range for teachers' pay in most OECD countries.

3 Data

The data we use in this study are based on standardized national tests and questionnaires (Growth and Effectiveness Measures for Schools, or GEMS; Meitzav in Hebrew) administered in 2005–2012.[14] The GEMS data that we use include fifth graders' test scores in native language skills (Hebrew or Arabic), math, science, and English, along with questionnaires polling fifth- and sixth-graders on their school learning environment, and questionnaires distributed to all teachers and principals. The available students' questionnaires date back to 2007, while those distributed to teachers and principals date only to 2009. [15] The GEMS data are described in more detail in Online Appendix A.

We linked the GEMS data to information available in the administrative records of the Israeli Ministry of Education for the entire elementary school student population in Israel, as well as for the teachers and schools. The student records include demographics that we use to construct all the student background measures: gender, parents' education, number of siblings, country of birth, and parents' country of origin. School records include information on enrollment, education stream (religious, etc.), school district, and the school's index of Socio-Economic Status (SES).[16] Teacher records include age, seniority (tenure), academic degree, and weekly working hours. Our dataset also includes the Ministry of Education's record of the year the school joined the reform.

Our analysis focuses on fifth- and sixth-grade pupils in the Jewish state school system, including both secular and religious schools.[17] The analysis excludes schools that joined

[14] Our sample ends in 2012 since by that time nearly all schools were already operating under the reform's rules. In addition, GEMS grades at the school level have been publicly available since 2012, following a decision by Israel's Supreme Court. This creates incentives for principals to manipulate grades (e.g., by selecting which students will be tested), and hence can lower the accuracy of the GEMS dataset from 2013 onward.

[15] Students' questionnaires from 2005 and 2006 were also available, but the questions were worded very differently, making it difficult to compare between these and later years.

[16] The school SES index is an average of the SES scores for its students. Student SES is a weighted average of values assigned to parents' schooling and income, economic status, immigrant status and former nationality, and the school's location (urban or peripheral). The index ranges from 1 to 10, with 1 representing the highest socioeconomic level. Schools with more disadvantaged students (high SES scores) receive more funding per student. Our data showed only the school SES (i.e., the average of the student indices).

[17] The sample is restricted to Jewish state schools (columns 1 and 2 in Table 1) that follow the same national curriculum and participate in the GEMS national testing. For these reasons, we exclude Arab schools (column 3 in Table 1) and private Jewish schools.

the reform during its first year of implementation (2008), schools that were part of a previous reform, and schools that closed before or opened after 2008 (approximately 30% of the schools in total).[18] We also exclude 12 schools that did not implement the reform by 2012.

The fifth-graders' GEMS test scores are available at the student level; hence the unit of observation for our test score analyses is the student. The raw test scores use a 1–100 scale that we converted into z-scores by year and subject. The test scores analysis file includes scores for 196,549 fifth-grade students from 877 schools. About 115,000 students were tested in math and language and about 114,000 students were tested in science and English. The test scores analysis covers data from 2005 through 2012. Because the sample of GEMS schools changes from year to year, the data structure is a repeated cross-section. Also, from 2007 onwards, schools that tested students in math and language in a given year were not the same schools that tested in science and English in the same year. Hence, not all students tested in all subjects.[19]

To protect students' privacy, the fifth- and sixth-grade GEMS questionnaires are not identified by student. Hence, we aggregate the data to the class level by calculating the average value of the student's answers for each item in the questionnaires, and then standardizing the class averages by cohort (year and grade). Twenty-eight items were used to measure school learning environment outcomes. To enable the testing of multiple hypotheses directly and to facilitate more precise conclusions about the reform effect, we group the statements into six indices: personal relations between students and teachers; teachers' efforts to help students; general satisfaction with the school; students' misbehavior in class; physical bullying; and social bullying. Each of the six indices is a simple (within-class) mean of the relevant standardized variables. Online Appendix A provides

[18] As mentioned in section 2, schools that implemented the reform in 2008 differed significantly from schools that implemented the reform afterward. We exclude schools that took part in a previous reform because it would not be possible to separate the effects of the two reforms.

[19] Nevertheless, the selection which schools would participate in the GEMS in a given year, along with their assigned subjects, was determined by the Ministry of Education in 2006, prior to the introduction of the reform. Schools were obligated to partake in these exams as outlined, rendering any concerns about selection in our GEMS data that correlate with the timing of the reform implementation irrelevant.

a detailed description of the GEMS student questionnaires and the construction of the six indicators. The school learning environment data analysis file includes 11,008 fifth- and sixth-grade cohorts from 877 schools. The school learning environment analysis covers data from 2007 through 2012. The data structure is a repeated cross-section, where each school is sampled once every two years.

We supplement our dataset with administrative data on teachers and principals. The administrative data includes demographic information on the teacher's role in the school (teaching subject, position, rank), and information on tenure and working hours. We compute averages by school and year for the following variables: total tenure, age, percentage of teachers with more than 30 years of experience, percentage of teachers with under 5 years of experience, and percentage of teachers with an academic degree. We also calculate for each school in each year the management tenure of the principal, and whether he or she was new at the school (a dummy variable which equals one if this was the principal's first year at the school, and zero otherwise).

Table 2 presents descriptive statistics for students, classes, teachers, and schools in the estimation samples. Means and standard deviations are reported for the fifth-grade test scores sample in columns 1 and 2 and for the fifth- and sixth-grade school learning environment sample in columns 3 and 4. The two samples have similar characteristics (similar characteristics also appear when looking at sub-samples by subject in the fifth-grade test scores sample). Given that the school learning environment sample is representative of all elementary schools and includes half of them each year and all of them over two years, the similarity in characteristics between the two samples suggests that the test scores sample is also a representative sample. Panel A presents the schools' characteristics, and panel B presents the characteristics of students (columns 1 and 2) or classes (columns 3 and 4) for the test-scores sample and school learning environment sample, respectively. Panel C presents the outcome variables. As can be seen, the average class size was 27, and there were on average about 20 teachers per school. Teachers worked on average 24–25 hours per week before the reform. The average management tenure of principals in our

Table 2: Descriptive Statistics

	Test Scores Sample		School Learning Environment Sample	
	Mean	**S.D.**	**Mean**	**S.D.**
	(1)	(2)	(3)	(4)
Panel A: School Characteristics				
Grade enrollment	58.37	28.04	59.55	27.95
Class enrollment	27.09	6.059	27.03	5.548
SES index	4.746	2.188	4.696	2.200
Periphery indicator	0.248	0.432	0.249	0.432
Number of teachers	21.71	14.16	18.87	14.35
Mean tenure – principals	10.12	6.445	10.20	6.598
% New principals (1yr)	0.047	0.044	0.047	0.045
Mean tenure – teachers	17.39	4.416	17.65	4.675
% New teachers (<5yr)	0.107	0.012	0.107	0.014
% Veteran teachers (>30yr)	0.130	0.020	0.139	0.024
% Academic degree	0.809	0.140	0.830	0.144
Religious school indicator	0.333	0.471	0.335	0.472
Average teachers' working hours	23.97	5.714	25.39	5.814
Number of schools	877		877	
Panel B: Student/Class Characteristics				
Male indicator	0.497	0.500	0.503	0.225
Father's years of education	12.18	4.906	12.05	2.688
Mother's years of education	12.67	4.397	12.59	2.634
Number of siblings	1.735	1.242	1.740	0.711
Born in Israel	0.938	0.241	0.932	0.087
Israeli ethnicity	0.619	0.486	0.611	0.194
Former USSR ethnicity	0.149	0.356	0.152	0.172
Ethiopian ethnicity	0.033	0.177	0.040	0.108
Asia-Africa ethnicity	0.091	0.288	0.087	0.078
Europe-America ethnicity	0.108	0.310	0.110	0.111
Number of students/classes	196,549		11,008	
Panel C: Outcome Variables				
I. Test Score Outcomes				
Math score [N=115,360]	65.79	21.60		
Language score [N=114,860]	73.73	17.62		
Science score [N=114,382]	70.19	19.47		
English score [N=112,928]	74.71	21.43		
II. School Learning Environment Outcomes				
Student–teacher relations [N=11,008]			-0.217	0.774
Teachers' efforts to help students [N=10,995]			-0.167	0.839
General school satisfaction [N= 11,008]			-0.036	0.840
Students' misbehavior in class [N=11,008]			0.147	0.783
Has suffered from physical bullying [N=11,004]			0.153	0.077
Has suffered from social bullying [N=11,008]			0.146	0.077

Notes: The table reports descriptive statistics for the sample of fifth and sixth-grade students in Jewish state elementary schools who participated in the GEMS tests in 2005-2012. Means and standard deviations for class-level data are computed using one observation per class; Means and standard deviations for school-level data are computed using one observation per school; Means and standard deviations for student-level data are computed using one observation per student.

sample was 10 years, and about 5 percent of the principals were in their first year at the school. Teachers' average tenure in the profession was about 17 years, more than 80% had an academic degree, 10% were new teachers with less than 5 years of experience, and 13% were veterans with more than 30 years of experience. A third of the schools were religious, and a quarter were in the periphery. Demographic data show that 94 percent of the students were Israeli-born. Almost 40% were the children of immigrants, with the largest group of those coming from the former Soviet Union (about 15 percent of the total sample). Parents of the children in the sample had on average 12 years of education.

4 Empirical Strategy

4.1 Estimation

The gradual implementation of the NH reform allows us to employ an event study methodology, or a generalized difference-in-differences design with staggered adoption, to detect the effect of the reform on students' outcomes. As shown by Lafortune et al. (2018b) educational reforms may affect students' outcomes - in particular academic achievement outcomes - slowly and gradually rather than in the same year, as a student's performance in year t likely depends in part on the quality of the schooling she received in prior years. Hence, given sufficient data for the years preceding the reform, we use an event study specification to identify the dynamic effects in the years after the reform. We also use the standard two-way fixed effects (TWFE) difference-in-differences estimator, as well as a parametric DiD proposed by Lafortune et al. (2018b) that both captures immediate effects and allows us to approximate dynamic effects that develop gradually over time.

Since we have data on test scores from 2005, we study the effect of the NH reform on students' achievements in math, language, science, and English by estimating the following event study specification separately for each subject:

$$y_{ist} = \sum_{j \neq -1} \beta_j \mathbb{1}[t = t_s^* + j] + \alpha_s + \delta_t + \lambda S_{st} + \gamma X_{ist} + \rho_{st} + u_{ist} \tag{1}$$

where y_{ist} is the outcome of interest - test scores of student i in school s in year t; t_s^* is the year in which school s implemented the reform; and j denotes the year relative to the reform implementation. We denote $j = 0$ as the first full year of the reform. We truncate j at -6 since we have data from 2005, and only 1% of observations were observed 7 years before their school reformed. We omit the event time dummy at $j = -1$, implying that the event time coefficients β_j measure the impact of the reform j years relative to the year just before the reform.[20] α_s and δ_t are school and year fixed effects, respectively. School fixed effects control for any systematic differences in our outcomes across schools. Year-fixed effects flexibly control for any overall trends in our main outcome variables that are common across schools. We also include S_{st} which is a vector of time-varying controls comprising characteristics of school s at time t, including SES index, SES index interaction with post-2008,[21] and log enrollment; and X_{ist}, which is a vector of characteristics of student i in school s in year t. Student characteristics include a gender dummy, both parents' years of schooling, the number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. Since district managers are responsible in part for the timing of the implementation of the reform in each school in their district we also include district-by-year fixed effects denoted by ρ_{st}.[22]

The school learning environment dataset is available only from 2007, and we observe each school only once in the pre-reform years. Hence, we estimate a standard difference-in-differences specification for both test score outcomes and learning environment outcomes as the following:

$$y_{ist} = \beta_1 Reform_{st} + \alpha_s + \delta_t + \lambda S_{st} + \gamma X_{ist} + \rho_{st} + u_{ist} \tag{2}$$

[20] As noted in Borusyak et al. (2021), when there are no never-treated units in the sample, two relative time indicators need to be omitted to avoid multicollinearity. Since we do not have never-treated schools in our sample we also drop the most negative relative time indicator ($t = -6$), so that the coefficients for the relative time indicators can be viewed as the mean differences from the average value of the outcomes in two specific relative periods before treatment.

[21] The SES index formula was changed in 2008.

[22] In Section 5 we also show results from specifications that do not include district-by-year fixed effects or include district linear time trend instead.

where i denotes the unit of analysis which, depending on the outcome, at the student level (test scores) or class level (learning environment outcomes). The variable $Reform_{st}$ is a dummy variable that indicates whether school s participated in the reform in year t and the vector X_ist is comprised of student characteristics (or class averages when the analysis is at the class level) and also includes dummy for sixth grade (when the outcome variables are the learning environment measures). The coefficient estimate β_1 represents the change in the outcome following the implementation of the reform. In implementing a staggered difference-in-differences analysis we follow the analytic approach used by Lafortune et al. (2018b) to accommodate both an immediate, i.e. same-year, and a gradual effect. This approach is presented in equation (3):

$$y_{ist} = \beta_1 Reform_{st} + \beta_2 Reform_{st} * Reform\_trend_{st} + \alpha_s + \delta_t + \lambda S_{st} + \gamma X_{ist} + \rho_{st} + u_{ist} \tag{3}$$

where $Reform\_trend_{st}$ is a trend variable that measures the years between year t and the year when the reform was implemented for each school s; it can take negative values before the implementation.[23] Here β_2 captures the delayed effects of the reform and represents the annual change in outcomes in school s from the implementation year, relative to the same school before the reform. Throughout our analyses, we use standard errors clustered at the school level.

4.2 Identifying Assumptions

To interpret the results of our analysis as a causal treatment effect of the NH reform on student outcomes, we rely on three key identifying assumptions: no anticipatory behavior prior to treatment, parallel trends in baseline outcomes, and treatment effect homogeneity (Sun and Abraham, 2021, Borusyak et al., 2021). Under the no anticipatory effects assumption, we assume that units do not change their behavior in anticipation of the treatment. Probably the main channel through which anticipation of the reform might affect

[23] Lafortune et al. (2018b) also include $Reform\_trend_{st}$ as a control, and suggest it represents a falsification test. As our treatment is at the school level and we include school-fixed effects and year-fixed effects, this term is omitted. The leads in our event study specification can represent a falsification test instead.

our results is via teachers. However, since the NH was a whole-system reform, teachers could not move between schools as a means to avoid it. Avoiding the reform was possible through early retirement, but we do not find such an effect in our sample, as we will show in Section 6.2. In addition, we consider it unlikely that teachers would deviate from their teaching practices in year t in anticipation of a reform that would change their work environment in year $t+1$ and (RAMA, 2008 find no evidence of such an effect). As for the students, since parents are unable to choose the school their child goes to without changing their address and since all schools eventually enter the reform, we do not expect an anticipatory response from parents.

The parallel trends assumption is the idea that absent the reform, the difference in outcomes (in our case, student test scores and the school learning environment) would be similar across all units and all periods conditional on the set of controls (school time-varying controls and student characteristics), and unit and time fixed effects. This assumption needs to be tested since the reform was not applied randomly, and some schools made an active choice to join the reform early. Even if early entrants to the reform differ from late entrants in some attributes, the existence of a parallel trend would alleviate concerns that this difference is correlated with the treatment timing. In our results, we will assess the validity of these assumptions visually by studying the dynamics of the event study coefficients in the pre-reform periods. As we will show, for the test score outcome variables, as well as for teachers' characteristics, these coefficients do not differ significantly from zero. Furthermore, in Figure B1 in Online Appendix B we present findings from estimating specifications similar to equations (1) with students' and schools' time-varying predetermined characteristics as the outcome variables.[24] The results in Figure B1 do not reveal any differential pre-trend in students' and schools' observables, thereby lending additional support to our assumption of parallel trends.

Recent econometric studies have highlighted that event study coefficients may be bi-

[24] We do not include teachers' predetermined characteristics, as we anticipate that the reform may also affect teachers and could alter their composition within the school. We discuss this further in Section 6.2. We also do not include SES index as an outcome variable because its formula was changed in 2008, at the same time the reform was launched.

ased if there is heterogeneity in treatment effects between groups of units treated at different times (see, among others De Chaisemartin and d'Haultfoeuille, 2020, Goodman-Bacon, 2021, Callaway and Sant'Anna, 2020 and Sun and Abraham, 2021). In these cases, each event time coefficient may be "contaminated" with effects from other cohorts. Specifically, Sun and Abraham (2021) show that the coefficients are linear combinations of cohort-specific effects from the given relative time period and other relative periods. The presence of heterogeneous treatment effects can lead to negative weights, potentially causing the estimated treatment effect to be negative even if the true average treatment effects are all positive (De Chaisemartin and d'Haultfoeuille, 2020).[25] We do not expect this issue to affect our results since this problem arises mainly when estimating long-run effects, while we have a short panel and focus on the short-run effect (Borusyak et al., 2021). Nevertheless, given the potential for biased estimates, we also estimate the event study specification (equation (1)) using an alternative estimator proposed by Sun and Abraham (2021) (SA) that allows for heterogeneous treatment effects. We show that using this alternative estimator generates consistent results, which further suggests that heterogeneity in treatment effects is unlikely to be an important concern in our setting.

5 Results

In this section, we present estimates for the effect of the NH reform on students' performance and indices of the school learning environment. We start by showing the effect on test scores and then turn to the effect on the school learning environment.

5.1 The Effect of the NH Reform on Test Scores

Figure 1 presents the effect of the NH reform on test scores in the four GEMS subjects: math, language, science, and English (panels a, b, c, and d respectively). The dark-blue diamonds and lines represent the point estimates and the 90 percent confidence intervals,

[25] As shown by Goodman-Bacon (2021) the inclusion of a control group can partially help to alleviate this issue. Unfortunately, as almost all schools were reformed by 2012, we are unable to include never-treated schools as a control group.

respectively, from estimating the event study specification (equation (1)) using a standard two-way fixed effects estimator (i.e. TWFE). The light-blue circles and lines represent the point estimates and the 90 percent confidence intervals, respectively, from estimating the event study specification (equation (1)) using the interaction-weighted estimator proposed by Sun and Abraham (2021) (i.e, SA), where the schools who joined the reform in the last year (2012) serve as a control group.

The results in Figure 1 show that the reform affects only math and English test scores and has no effect on language and science test scores. Moreover, in Math and English, the effect is not immediate, as it only begins in the second year of the reform. Math test scores increase by about 0.15–0.2 standard deviations after two years of the reform ($t > 0$) where the SA estimates are slightly weaker. The effect on English test scores takes more time to materialize, with a moderate and only marginally statistically significant increase of about 0.1 standard deviations in the second year of the reform; while in the fourth year, there is a large and statistically significant increase in test scores of about 0.2 standard deviations ($t = 3$). Moreover, there are no indications of differential trends in test scores across any subject before the implementation of the reform which supports the parallel trends assumption.

The TWFE and SA methods provide qualitatively similar results, hence we estimate equations (2) and (3) using TWFE since heterogeneity in treatment effects is unlikely to be an important concern in our setting. Table 3 reports the estimates of equation (2) in Panel A and equation (3) in Panel B for the standardized test scores in the four GEMS subjects (math, language, science, and English). The effects are estimated using specifications that control for both students and school characteristics (columns 1, 4, 7, and 10), specifications that control for both students and school characteristics and include district-by-year fixed effects (columns 2, 5, 8, and 11) and specifications that include district-specific linear time trends instead of district by year fixed effects (columns 3, 6, 9 and 12). All specifications include school and year-fixed effects as well as student and school characteristics.

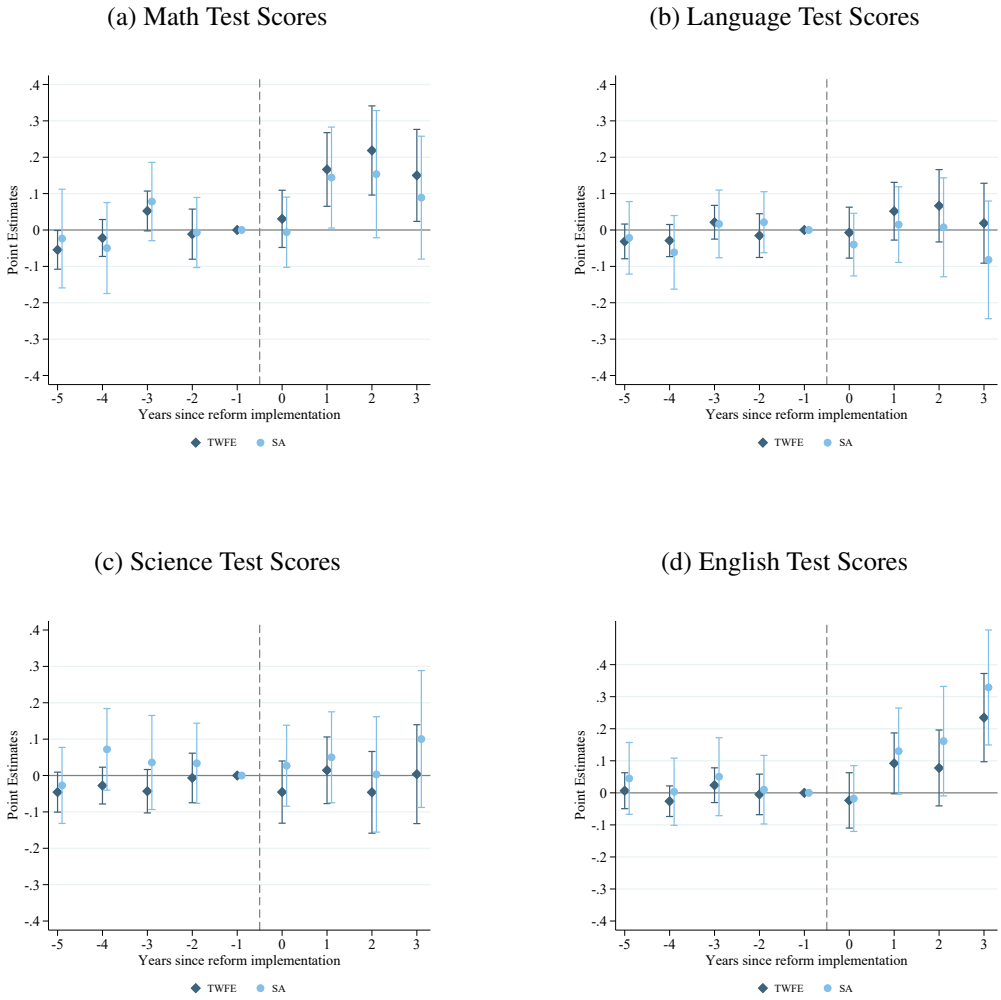
Table 3: Effect of the NH Reform on Test Scores - DiD Estimation

	Math			Language			Science			English		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A:												
Post-reform	0.060	0.061	0.064	0.022	0.017	0.024	-0.009	-0.021	0.002	0.015	0.022	0.200
	(0.039)	(0.039)	(0.039)	(0.032)	(0.033)	(0.033)	(0.038)	(0.039)	(0.039)	(0.040)	(0.042)	(0.040)
Panel B:												
Post-reform	0.046	0.043	0.049	0.021	0.012	0.022	-0.012	-0.023	0.001	-0.008	-0.002	0.0005
	(0.039)	(0.039)	(0.039)	(0.033)	(0.034)	(0.034)	(0.039)	(0.040)	(0.040)	(0.041)	(0.043)	(0.041)
Post-reform Trend	0.035**	0.043**	0.039**	0.002	0.010	0.005	0.009	0.005	0.002	0.064***	0.067***	0.057***
	(0.018)	(0.020)	(0.019)	(0.016)	(0.017)	(0.016)	(0.022)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Student Controls	X	X	X	X	X	X	X	X	X	X	X	X
School Controls	X	X	X	X	X	X	X	X	X	X	X	X
District by Year FE		X			X			X			X	
District Linear Time Trend			X			X			X			X
Students		115,360			114,860			114,382			112,928	
Schools		877			877			872			870	

Notes: The table reports estimates of parametric difference-in-differences models corresponding to equation (2) in panel A and equation (3) in panel B. The dependent variable is the student's standardized test score by year in math (columns 1–3), language (columns 4–6), science (columns 7–9), and English (columns 10–12). All specifications include school and year-fixed effects. Student controls include a gender indicator, both parents' years of schooling, the number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. School controls include SES index, the interaction between the SES index and a dummy for the post-reform period, and the log of school enrollment. Standard errors are clustered at the school level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 1: Effect of the NH Reform on Test Scores - Event Study Estimation



Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variable is the student's standardized test score by year in math (panel a), language (b), science (c), and English (d). The dark-blue diamonds indicate the coefficients from estimating equation (1) using TWFE and the light-blue circles indicate the coefficients from estimating equation (1) using SA. All specifications include the full set of event time dummies, school-fixed effects, year-fixed effects, district-by-year fixed effects, and controls for student and time-varying school characteristics. Student characteristics include a gender dummy, both parents' years of schooling, the number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. Time-varying school characteristics include SES index, the interaction between the SES index and a dummy for the post-reform period, and the log of school enrollment. The sample includes fifth-grade students from 877 Jewish (Hebrew-speaking) state elementary schools that participated in the GEMS tests between 2005 and 2012. Standard errors are clustered at the school level.

The estimated effects of the post-reform dummies in panel A are not statistically different from zero. Moving to panel B, which adds the post-reform trends, we see that the reform has positive and statistically significant effects only on math and English test scores, with no effect on test scores in language and science. Moreover, the effects in math and English are gradual rather than immediate. These results are equivalent to the results in Figure 1. The estimates in Table 3 are similar across all specifications. Accord-

ing to the estimates from our preferred specification that includes district-by-year fixed effects, math and English test scores rose following the reform by an average of 0.043 and 0.067 standard deviations per year, respectively (columns 2 and 11), adding up to an average increase of 0.13 and 0.2 standard deviations three years after the reform was implemented. In section 6 we explore possible explanations for why the NH reform affects students' performance in this manner, and why the effect is seen only in math and English and not in language and science. The results presented in Table 3 essentially give an average post-reform effect for the year effects presented in Figure 1.

A potential concern with our empirical approach is that the impacts may be driven by the earliest schools to participate, which could bias the results if these schools were particularly eager to participate in the reform, or different in some other way. To test whether our results are driven by a specific cohort of schools that implemented the reform in a given year, Figure B2 in Online Appendix B displays results from estimating equation (1) using TWFE, while excluding each time a different cohort of schools based on the year of reform implementation. The figure shows similar results as Figure 1. Specifically, the effect of the reform on math test scores begins in the second year of implementation, regardless of the implementation year. Moreover, the impact on English test scores is not driven by the earliest participating schools, as the score increase is evident two and three years after implementation even when excluding schools that reformed in 2009. Figure B1 in Online Appendix B also provides evidence that the timing of the reform does not correlate with predetermined student characteristics and time-varying school characteristics.

5.2 The Effect of the NH Reform on the School Learning Environment

Tables 4 and 5 report the estimated effects of the NH reform on our six indices of the school learning environment in the same format as Table 3, using the same specifications but with the class-level dataset. The estimates in panel A show that classes in reformed schools immediately scored higher on the indices of student–teacher relations and teach-

ers' efforts to help students. The point estimates are 0.092 for student–teacher relations and 0.117 for teachers' efforts (Table 4, columns 2 and 5, respectively). Also, classes in reformed schools report lower levels of student misbehavior in class, with a point estimate of -0.103 (Table 5, column 2). These results are not sensitive to the inclusion of and the functional form of district-specific time trends. Other school learning environment indices do not seem to be affected by the NH reform. These include general school satisfaction (Table 4, columns 7–9) and physical and social bullying (Table 5 columns 4–9).

The results of panel B are similar to those of panel A. It is important to underscore that the effects of the NH reform on the school learning environment materialize immediately and exhibit no time trend. This is in contrast to the effect of the reform on test scores, which is only observable two or three years after the reform was implemented. Thus, while test scores are the result of cumulative and continuous learning efforts, reported measures of the school learning environment can (and do) change immediately once changes to the learning environment are made. Unfortunately, we observe the measures of the school learning environment only once for each school in the pre-reform period, hence, we can not test for pre-reform time trends in these measures. Nevertheless, the absence of a pre-reform trend in test scores and teachers' characteristics, as shown in Section 6.2, suggests that the parallel trend assumption is not violated in the measures of the school learning environment as well.

6 Potential Channels for the Effect of the Reform

The NH reform brought changes in several educational inputs which in turn brought positive effects on math and English test scores and the school learning environment. In this section we discuss the potential contributions of the main elements of the reform to the observed effects. We discuss two possible channels: the direct effect on students through small-group learning activities, and the indirect effects on students stemming from the effect of the reform on teachers through the change in their working conditions. The other elements of the reform, namely incentives for professional development and pro-

Table 4: Effect of the NH Reform on the School Learning Environment - DiD Estimation (Part 1)

	Student–teacher relations			Teachers' efforts to help students			General school satisfaction		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A:									
Post-reform	0.092***	0.092***	0.097***	0.111***	0.117***	0.117***	0.026	0.024	0.031
	(0.033)	(0.033)	(0.033)	(0.034)	(0.034)	(0.034)	(0.038)	(0.038)	
Panel B:									
Post-reform	0.103***	0.102***	0.108***	0.117***	0.122***	0.122***	0.040	0.037	0.044
	(0.034)	(0.034)	(0.034)	(0.035)	(0.035)	(0.035)	(0.038)	(0.039)	(0.039)
Post-reform trend	0.034	0.031	0.033	0.019	0.016	0.017	0.043	0.040	0.039
	(0.021)	(0.021)	(0.022)	(0.022)	(0.022)	(0.022)	(0.026)	(0.027)	(0.027)
Student Controls	X	X	X	X	X	X	X	X	X
School Controls	X	X	X	X	X	X	X	X	X
District by Year FE		X			X			X	
District Linear Time Trend			X			X			X
Classes		11,005			11,005			11,005	
Schools		877			877			877	

Notes: The table reports estimates of parametric difference-in-differences models corresponding to equation (2) in panel A and equation (3) in panel B. The dependent variables are standardized scores for the class average in each of the school learning environment indicators. All specifications include school and year-fixed effects. Student controls are class averages of boys, both parents' years of schooling, the number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. School controls include SES index, the interaction between the SES index and a dummy for the post-reform period, and the log of school enrollment. Standard errors are clustered at the school level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Effect of the NH Reform on the School Learning Environment - DiD Estimation (Part 2)

	Student misbehavior in class			Suffered from physical bullying			Suffered from social bullying		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A:									
Post-reform	-0.104***	-0.103***	-0.110***	-0.004	-0.004	-0.004	-0.003	-0.004	-0.004
	(0.038)	(0.038)	(0.038)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Panel B:									
Post-reform	-0.112***	-0.114***	-0.121***	-0.004	-0.004	-0.005	-0.003	-0.004	-0.004
	(0.039)	(0.039)	(0.039)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Post-reform Trend	-0.026	-0.034	-0.032	-0.002	-0.002	-0.002	-0.0004	-0.0006	-0.0009
	(0.024)	(0.025)	(0.025)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Student Controls	X	X	X	X	X	X	X	X	X
School Controls	X	X	X	X	X	X	X	X	X
District by Year FE		X			X			X	
District Linear Time Trend			X			X			X
Classes		11,005			11,005			11,005	
Schools		877			877			877	

Notes: The table reports estimates of parametric difference-in-differences models corresponding to equation (2) in panel A and equation (3) in panel B. The dependent variables are standardized scores for the class average in each of the school learning environment indicators. All specifications include school and year-fixed effects. Student controls are class averages of boys, both parents' years of schooling, the number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. School controls include SES index, the interaction between the SES index and a dummy for the post-reform period, and the log of school enrollment. Standard errors are clustered at the school level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

motion, probably have a long-term rather than a short-term effect, since both take time to materialize.

6.1 Small-Group Learning

According to RAMA's evaluation reports (RAMA, 2012a, RAMA, 2012b), teachers reported that they devoted the majority of the small-group learning hours to students with learning difficulties and students with average achievement (on average 42 percent and 34 percent of the small-group learning hours, respectively). Ten percent of the small-group learning hours were devoted to outstanding students, and the rest to heterogeneous or other groups.

RAMA's evaluation report further documents that the small learning groups comprised 4.5 students on average, and about 63 percent of the students in schools under the reform attended small-group learning classes at least once a year. Students' participation in the groups lasted five months (23 weeks) on average. Teachers determined when students' participation in small-group learning ended (usually when the student's performance improved to the teacher's satisfaction). Teachers also reported that the majority of small-group hours (about 80 percent) were allocated to the reinforcement of core subjects, especially reading and writing in the student's mother tongue (i.e., Hebrew in Jewish schools) and mathematics.

Given that the group learning activities are aimed largely at students with initial low and average achievement, it is of interest to examine whether such students benefited from the reform more than others. As we are unable to observe students' previous achievement, we analyze heterogeneous treatment effects using parents' education as a proxy for students' achievement. Parental education is highly correlated with student achievement; hence, having a poorly educated parent is a proxy for a poorly achieving student (RAMA, 2012a, RAMA, 2012b). In our student-level data, we calculate the median value of the parents' schooling for each class. Then we classify each student as having a low achievement level if his or her parents's schooling is below the class's median value. We estimate

equation (3) with students' test scores as an outcome variable and include an indicator for students with low-educated parents and the interaction of this indicator with the post-reform and the post-reform trend variables to allow a different effect of the reform on students with low-educated parents.

Table 6 reports the estimated effects of the NH reform using the aforementioned estimation in the same format as Table 3. The results in Table 6 show that the positive effect of the reform on math test scores is driven by its effect on students whose parents' sum of schooling is below the class median. It also shows that students with low-educated parents improve their language test scores due to the reform, though the estimates are smaller compared to the estimates for math and statistically significant only at 10% significant level. Math and language test scores among presumed low-achievement students rose by an average of 0.04 and 0.024 standard deviations, respectively, compared to the other students. It is apparent from these results that, in both math and language, the students who came from homes characterized by lower education levels, and who were therefore likely to have lower achievement levels in school, improved their test scores to a greater extent. Because lower-achieving students were more likely to be included in small-group learning, which in turn focused mainly on reinforcement of core subjects such as language and math than the other subjects, the findings provide suggestive evidence that the effect on math and language test scores that we observe in our main results probably traces to the small-group learning format.

In contrast, the effect of the reform on science test scores is insignificant for all students, and the effect of the reform on English test scores is very large, positive, and statistically significant for all students. Only 9 percent and 5 percent of small-group learning hours were devoted to English and science during the study period, respectively (RAMA, 2012a, RAMA, 2012b). As we do not observe differential effects by parents' schooling, we may assume that the effects on English test scores are not driven by the small-group learning classes but rather by some other component of the reform, such as the effect of the reform on English teachers and their teaching practice in regular classes. In the next

Table 6: Effect of the NH Reform on Test Scores by Parents' Schooling - DiD Estimation

	Math			Language			Science			English		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Post-reform	0.057	0.054	0.060	0.024	0.015	0.015	-0.013	-0.026	-0.001	-0.004	0.001	0.003
	(0.039)	(0.038)	(0.039)	(0.034)	(0.034)	(0.034)	(0.040)	(0.040)	(0.040)	(0.042)	(0.044)	(0.042)
Post-reform*	-0.028	-0.032	-0.028	-0.004	-0.006	-0.006	0.010	0.010	0.014	-0.003	-0.004	-0.001
Low-educated Parents	(0.023)	(0.023)	(0.023)	(0.022)	(0.022)	(0.022)	(0.023)	(0.023)	(0.023)	(0.021)	(0.021)	(0.021)
Post-reform Trend	0.021	0.027	0.024	-0.007	0.0004	0.0004	0.008	0.005	0.002	0.058***	0.061***	0.052**
	(0.018)	(0.020)	(0.019)	(0.017)	(0.017)	(0.017)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Post-reform Trend*	0.038***	0.040***	0.038***	0.025*	0.024*	0.024*	0.002	0.0008	0.00005	0.018	0.017	0.017
Low-educated Parents	(0.014)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.015)	(0.015)	(0.015)	(0.013)	(0.013)	(0.013)
Student Controls	X	X	X	X	X	X	X	X	X	X	X	X
School Controls	X	X	X	X	X	X	X	X	X	X	X	X
District by Year FE		X			X			X			X	
District Linear Time Trend			X			X			X			X
Students		115,360			114,860			114,382			112,928	
Schools		877			877			872			870	

Notes: The table reports estimates of parametric difference-in-differences models corresponding to equation (3) with the following additional variables: an indicator for students with parents' schooling below the class's median value, an interaction of this indicator with the post-reform variable, and an interaction of this indicator with the post-reform trend variable. The dependent variable is the student's standardized test score by year in math (columns 1–3), language (columns 4–6), science (columns 7–9), and English (columns 10–12). All specifications include school and year-fixed effects. Student controls include a gender indicator, both parents' years of schooling, the number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. School controls include SES index, the interaction between the SES index and a dummy for the post-reform period, and the log of school enrollment. Standard errors are clustered at the school level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

section, we will try to explain why the reform might have affected some teachers, mainly English teachers, differently, and why it improved English achievement for all students.

6.2 Teachers' Characteristics and Working Conditions

One of the key aspects of the NH reform is its changes to teachers' working conditions, including increases in teachers' working hours, time spent in school, and monthly salary, and the restructuring of incentives and training. Thus, the reform does not necessarily reward teachers for students' short-term success. Rather, it creates a system that induces teachers to invest more time and effort in instructional activities, and that is designed to attract and retain more able and motivated teachers. We would expect these effects to change the quality of instruction in the state education system over the medium and long term, and, hence, not necessarily to manifest as improvements in students' achievement in the short run. These changes, however, may also have immediate effects on teaching quality among teachers already in the system. In this section, we examine whether teachers' characteristics and working conditions offer a potential explanation for our main results.

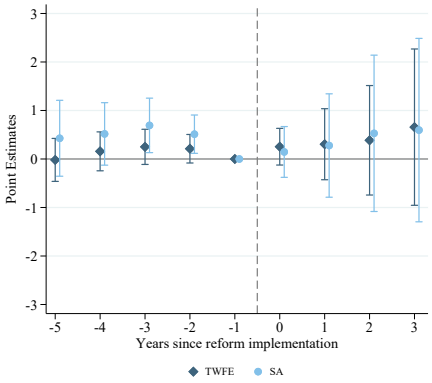
6.2.1 Composition of Teachers and Principals

We use administrative data on teachers and principals in order to analyze whether the reform changed the composition of either of these groups within schools. We focus on the principals' tenure and the appointment of new principals, as well as on school averages for the following variables: teachers' total tenure, percentage of teachers with more than 30 years of experience, percentage of teachers with less than 5 years of experience, and percentage of teachers with an academic degree. Since administrative data is available for each year of our sample period (2005–2012) for each school, we can use the event study specification depicted in equation (1) and test for preexisting trends. Instead of student controls, we include school averages of student characteristics.

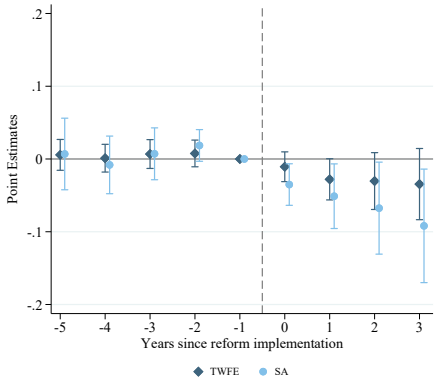
We start by analyzing the principals' tenure and whether principals are new at their schools. Figure 2 reports the results for the TWFE estimation (in dark-blue diamonds)

Figure 2: Effect of the NH Reform on Principals' Characteristics - Event Study Estimation

(a) Principals' Average Tenure



(b) Percentage of New Principals



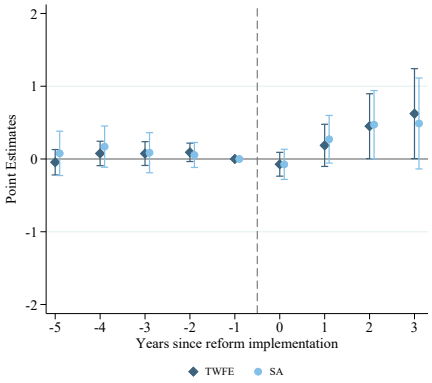
Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variable is principals' average tenure (panel a) and a dummy variable for whether principals are new at their school (b). The dark-blue diamonds indicate the coefficients from estimating equation (1) using TWFE and the light-blue circles indicate the coefficients from estimating equation (1) using the method of Sun and Abraham (2021) (i.e. SA). All specifications include the full set of event time dummies, school-fixed effects, year-fixed effects, district-by-year fixed effects, and controls for student and time-varying school characteristics. Student characteristics are school averages of both parents' years of schooling, number of siblings, percentage of boys, percentage born in Israel, and percentage for each ethnic origin. Time-varying school characteristics include SES index, the interaction between the SES index and a dummy for the post-reform period, and the log of school enrollment. The sample includes yearly observations of 877 elementary schools between 2005 and 2012. Standard errors are clustered at the school level.

and the results using SA estimation (in light-blue circles). The TWFE estimation provides statistically insignificant coefficients both for the years before and after the reform. However, after correcting for heterogeneous treatment effect, the findings in Figure 2 suggest that principals with a slightly higher tenure (of less than one year) tend to implement the reform in their school first, and after the reform implementation, principal turnover decreases. Nevertheless, the magnitudes of these changes are relatively small and the estimated effects of the NH reform on students' outcomes do not change when controlling for principals' tenure (results available upon request).

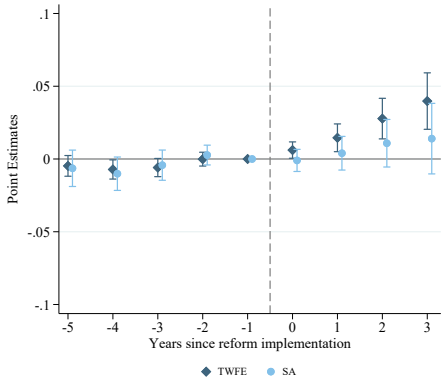
Figure 3 presents the estimates for the teachers' composition in school for the TWFE estimation (in dark-blue diamonds) and the SA estimation (in light-blue circles). Sub-figure (a) reports the results for the average tenure of teachers in the school, sub-figure (b) reports the results for the percentage of teachers with an academic degree in the school and the outcome variables in sub-figures (c), and (d) is the percentage of new teachers (less than 5 years) and veteran teachers (over 30 years) in school, respectively. The results

Figure 3: Effect of the NH Reform on Teachers' Characteristics - Event Study Estimation

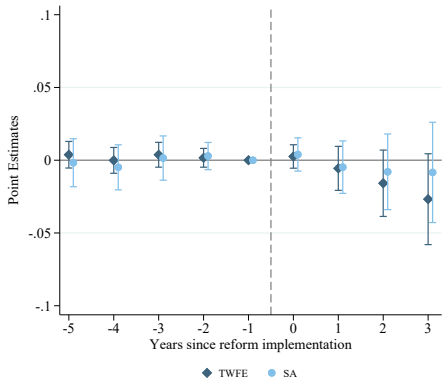
(a) Teachers' Average Tenure



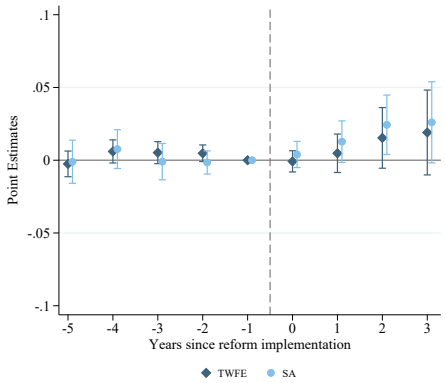
(b) Percentage of Teachers with Academic Degree



(c) Percentage of New Teachers (<5 Years)



(d) Percentage of Veteran Teachers (>30 Years)



Notes: Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variable is teachers' average tenure (panel a), the percentage of teachers with an academic degree (b), the percentage of new teachers, defined as under 5 years (c), and the percentage of veteran teachers, defined as more than 30 years (d). Regressions are estimated using OLS and include the full set of event time dummies, school fixed effects, year fixed effects, district by year fixed effects, and controls for student and time-varying school characteristics. Student characteristics are school averages of both parents' years of schooling, number of siblings, percentage of boys, percentage born in Israel, and percentage for each ethnic origin. Time-varying school characteristics include SES index, interaction between the SES index and a dummy for the post-reform period, and log of enrollment. The sample includes yearly observations of 877 elementary schools between 2005 and 2012. Standard errors are clustered at the school level.

show that teachers' tenure moderately increased after the reform, suggesting that turnover rates fell, although the coefficients are marginally statistically significant. The reform seems to have had an effect on the academic qualifications of teachers, as the percentage of teachers with an academic degree increased after the reform, while the TWFE estimates show an increase of 4 percentage points in the share of teachers with an academic

degree after three years of the reform. The increase in teachers with academic degrees is to be expected since one of the aims of the reform was to boost teachers' academic qualifications, and teachers were given incentives to complete their academic training. Hence, three years after the reform was implemented most teachers in the system obtained an academic degree (whether by completing one during these three years or by new teachers who were recruited and must have an academic degree). This can be viewed as a mechanical effect, which can explain why the SA estimates are insignificant. The post-reform changes in the percentages of new teachers and veteran teachers are not precise, but we do observe a marginal increasing trend in the proportion of veteran teachers. For all these variables, we observe no pre-event trend. This strengthens the validity of our identifying assumptions regarding anticipation and parallel trends discussed in Section 4.2.

6.2.2 Differences by Teaching Subject

Our main analysis indicates that the reform contributed to improvements in students' achievement in math and English. Section 6.1 provides suggestive evidence that small-group learning can explain the positive effect on math test scores, however, it does not provide an explanation for the improvement in English test scores. It might be that the improvement in English test scores was driven by the effect of the reform on teachers. To further explore this, we use the teachers' GEMS questionnaires to provide some information on the differences between teachers in different subjects.

Teachers in schools that participated in the GEMS were asked to answer questionnaires about their teaching experience. These questionnaires are available only from 2009 onward—i.e., the post-reform period. Thus, we cannot analyze these data in a way that would yield a causal interpretation. Nevertheless, these questionnaires can provide suggestive evidence regarding the heterogeneity of the effect of the reform on students, by subject.

We distinguish between teachers in the four different subjects (math, language, science, and English), and focus on teachers who teach these subjects in any grade, not

necessarily fifth or sixth. We selected seven questions on the teachers' questionnaire that required a yes-or-no answer: (1) was the teacher also a homeroom teacher; (2) was the teacher's subject area (math, language, science, or English) also the teacher's main specialization; (3) did the teacher also teach other subjects; (4) did the teacher have an academic degree; (5) did the teacher feel overloaded at work; and (6) was the teacher generally satisfied with the school and (7) work. For each question, we compute the percentage of teachers within the same teaching subject providing a positive answer within a given school. Table 7 presents the averages of these percentages and standard deviations by teaching subject and reform status for the subset of schools that we observe in the teachers' GEMS questionnaires.[26]

The summary statistics presented in Table 7 reveal differences between teachers of different subjects. Importantly, English teachers seem to differ most significantly from teachers of other subjects. First, we note that English teachers are much less likely to be homeroom teachers (20% of English teachers are homeroom teachers, compared to 88% of math teachers, 95% of Hebrew language teachers, and 60% of science teachers). Since homeroom teachers saw a greater increase in their workload compared to regular teachers (RAMA, 2012a, RAMA, 2012b), English teachers would have on average lower increases to their workloads compared to teachers of other subjects. Indeed, English teachers report lower levels of overload (55% compared to about 65% for teachers of other subjects). Second, English teachers are much more likely to specialize in their subject (63% for English teachers compared to 25% of math teachers, 13% of Hebrew language teachers, and 38% of science teachers), and to teach only their subject. However, they do not differ much from other teachers in school and work satisfaction. Summing up the results of Table 7, it seems that English teachers are more specialized and less overloaded compared to teachers of other subjects. These differences provide suggestive evidence which may explain the improved student achievement in English compared to other subjects.

[26]The GEMS teachers' questionnaire became available only in 2009, at a time when some schools had already transitioned to the reform rules. Therefore, we categorize outcomes from the teachers' questionnaire as 'before-reform' for schools that had not yet implemented the reform and 'after-reform' for schools that had, for each year between 2009 and 2012.

Table 7: Teacher Questionnaire Responses by Teaching Subject

	Math teachers		Language teachers		Science teachers		English teachers	
	Before-reform	After-reform	Before-reform	After-reform	Before-reform	After-reform	Before-reform	After-reform
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Homeroom teacher (1 = Yes, 0 = No)	0.879	0.880	0.943	0.946	0.608	0.616	0.201	0.195
	(0.326)	(0.325)	(0.231)	(0.227)	(0.488)	(0.486)	(0.401)	(0.396)
Main specialization in the subject (1 = Yes, 0 = No)	0.250	0.253	0.127	0.129	0.376	0.380	0.620	0.634
	(0.433)	(0.434)	(0.333)	(0.335)	(0.485)	(0.485)	(0.485)	(0.482)
Teaching different subjects (1 = Yes, 0 = No)	0.814	0.814	0.725	0.727	0.636	0.639	0.204	0.199
	(0.389)	(0.390)	(0.446)	(0.446)	(0.481)	(0.480)	(0.403)	(0.400)
Academic degree (1 = Yes, 0 = No)	0.873	0.882	0.871	0.880	0.868	0.875	0.872	0.882
	(0.333)	(0.323)	(0.335)	(0.325)	(0.339)	(0.330)	(0.334)	(0.322)
Overloaded at work (1 = Yes, 0 = No)	0.623	0.651	0.639	0.667	0.617	0.647	0.524	0.554
	(0.485)	(0.477)	(0.480)	(0.471)	(0.486)	(0.478)	(0.500)	(0.497)
School satisfaction (1 = Yes, 0 = No)	0.864	0.867	0.863	0.867	0.870	0.873	0.852	0.856
	(0.342)	(0.340)	(0.344)	(0.339)	(0.336)	(0.333)	(0.355)	(0.351)
Work satisfaction (1 = Yes, 0 = No)	0.937	0.937	0.936	0.938	0.917	0.918	0.897	0.899
	(0.244)	(0.243)	(0.246)	(0.242)	(0.275)	(0.275)	(0.304)	(0.301)
Teachers	13,632	10,483	16,919	12,952	5,751	4,419	3,184	2,429
Schools	414	414	414	414	414	414	414	414

Notes: The table reports means and standard deviations in parentheses for selected questions from the GEMS teachers' questionnaire by teaching subject and treatment status (before- and after-reform). The sample includes teachers from Jewish (Hebrew-speaking) state elementary schools that participated in the GEMS tests between 2009 and 2012 at least twice, once in their pre-reform period and once in the post-reform period.

6.2.3 Teachers Working Hours

An important component of the reform was the increase in teachers' paid working hours.[27] The administrative data in our dataset allow us to observe the working hours of each teacher in each school. An interesting test is whether the change in weekly working hours was similar for all teachers, as differences between teachers in different subject areas may offer a possible explanation for our results on test scores. Figure 4 shows the results for estimating equation (1) where the dependent variable in panel (a) is the average working hours per week of all teachers in the school, and in panel (b) it is the difference in working hours between homeroom teachers and English teachers at the school level, including school averages of student characteristics instead of student-level controls.[28]

The NH reform led to a substantial increase in teachers' working hours, with an average rise of nearly 8 weekly hours (panel (a)). A comparison between homeroom teachers and English teachers reveals that the rise in weekly working hours for English teachers was lower by approximately two hours (panel (b)). These findings support our earlier conclusion (from Table 7) that English teachers may have gained the most from the reform in terms of professional recognition and pay, without experiencing the heightened stress that often accompanies increased responsibilities at school. This presents another potential pathway through which the reform may have influenced student achievement in English.

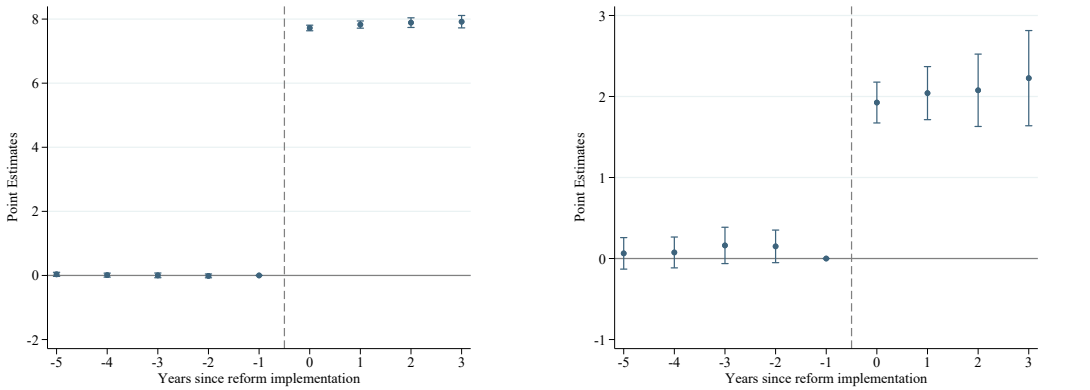
Unfortunately, we have no data on teachers' salaries, and therefore cannot estimate changes in teachers' hourly wages. According to Maagan (2015), teachers' hourly wages increased significantly following the reform. However, the hourly wages of principals and homeroom teachers actually decreased. Adding this to the evidence in Table 7, we conclude that in the short-run the reform worsened the working conditions of homeroom

[27] Before the reform teachers did many of their tasks at home (grading, class preparations etc.), without reporting these hours as work time. After the reform, the teachers were expected to perform these tasks in school and report these hours as work time.

[28] The increase in the working hours of teachers is a mechanical result of the reform - one of the attributes of the reform is the requirement that teachers increase their working hours. Therefore, we do not apply a correction for heterogeneous treatment effect by using Sun and Abraham (2021) estimation.

Figure 4: Effect of the NH Reform on Teachers' Working Hours - Event Study Estimation

(a) Average Working Hours (All Teachers) (b) Δ Working Hours (Homeroom - English Teachers)



Notes: Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variable is the average working hours of all teachers (panel a) and the difference in working hours between homeroom teachers and English teachers (b). Regressions are estimated using TWFE and include the full set of event time dummies, school-fixed effects, year-fixed effects, district-by-year fixed effects, and controls for student and time-varying school characteristics. Student characteristics are school averages of both parents' years of schooling, number of siblings, percentage of boys, percentage born in Israel, and percentage for each ethnic origin. Time-varying school characteristics include SES index, the interaction between the SES index and a dummy for the post-reform period, and the log of school enrollment. The sample includes yearly observations of 877 elementary schools between 2005 and 2012. Standard errors are clustered at the school level.

teachers, who usually teach math and language, and probably benefited the working conditions of English teachers. This conclusion may explain our findings on test scores.

Summing up the results of this section, we can see that improvements in student achievement in math are concentrated in lower-performing students. This matches the actual operation of the small-group learning sessions, which were mainly used to help such students in these subjects. In addition, we find differences between English teachers and teachers of other subjects. English teachers tend to be specialists rather than generalists. Beyond that, and partly as a result, they tend to feel less overloaded at work and to be better compensated in terms of their hourly salary. They also saw less of an increase in their actual workload following the reform. We believe that these differences contribute to their class performance, and hence also affect student achievement. Finally, we document an overall increase in teachers' working hours and small increases in teachers' and principals' average tenure. These improvements could be an additional explanation for the uniformly positive effect on the general school environment.

7 Conclusions

The NH reform was a nationwide educational reform that was gradually implemented in Israeli elementary schools starting in 2008. Its main aims were to improve academic achievement, narrow academic disparities and improve teacher working conditions. This study evaluates the short-term effects of the reform. The results, based on fifth-grade students' test scores, clearly indicate that the NH reform had a positive and significant short-term effect on student performance in math and English. The magnitudes of the estimated effects resemble those of other interventions that have been applied in the Israeli elementary education system, such as reducing class sizes (Angrist and Lavy, 1999) and strengthening teacher training (Angrist and Lavy, 2001), but exceed the estimated effects of providing more instruction time (Lavy, 2020) and using computer-aided instruction in the classroom (Angrist and Lavy, 2002).[29] Additionally, the reform helped to improve the school learning environment, as reflected in student questionnaires among fifth- and sixth-graders. The evidence suggest that the NH reform improved student–teacher relations, improved teachers' efforts to help students, and reduced misbehavior in class (as viewed by students).

This is the first paper that provides a detailed evaluation of the causal effects of the NH reform. In addition, we offer possible mechanisms by which this reform affected students' academic performance and the school learning environment. The evidence presented in this paper suggests that the positive effect of the reform on students' academic achievement probably traces to small-group learning and to the improvement in some subject teachers' working conditions. Another channel is the increased hours that teachers spend in school, which may improve the school learning environment through its indirect effect on students. It seems, however, that homeroom teachers experienced a worsening in their working conditions in the short run. This means that the effects of small-group learning

[29] Back-of-the-envelope calculations show that the cost of the reform per class is slightly higher than the cost of reducing class sizes or training teachers (Angrist and Lavy, 2001). The estimated cost of the reform is $1.4 billion annually (Bank of Israel, 2018). We note, however, that the estimated effects reported in this paper are short-term effects; the long-term effects might be greater.

may be greater than our estimates, which relate to the overall effect of the reform.

Our study may have broader implications that extend beyond its contribution to the Israeli education system. The NH reform includes intervention components commonly found in education systems worldwide. However, studies evaluating the effects of these components in a nationwide education reform context are rare. Notably, our pattern of results aligns with the extensive literature on U.S.-based whole-school reforms that aim to improve low-performing schools (Schueler et al., 2022). Therefore, our findings may be relevant for centralized policymakers in general. However, the effect of the NH reform on test scores appears larger than in U.S.-based whole school reforms studies. One possible explanation could be that Israel's education system is characterized by comparatively lower teacher salaries, larger class sizes, and lower per-student spending compared to the US and other OECD countries. Thus, our findings may be regarded as an upper bound for education systems with higher per-student spending.

The results of this study show a positive effect of the reform in the short run. Moreover, since one of the key aspects of the reform is the change in teachers' working conditions and the structure of incentives, NH may create a system that attracts and retains more able teachers and produces larger positive effects in the long run. Therefore, future research on the long-term effects of the reform and an in-depth cost–effectiveness analysis are clearly needed.

References

Angrist, J. and V. Lavy (2002). New evidence on classroom computers and pupil learning. *The Economic Journal 112*(482), 735–765.

Angrist, J. D. and V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics 114*(2), 533–575.

Angrist, J. D. and V. Lavy (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics 19*(2), 343–369.

Angrist, J. D., V. Lavy, J. Leder-Luis, and A. Shany (2019). Maimonides' rule redux. *American Economic Review: Insights 1*(3), 309–324.

Araujo, M. C., P. Carneiro, Y. Cruz-Aguayo, and N. Schady (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics 131*(3), 1415–1453.

Bank of Israel (2018). Bank of israel annual report 2018, chapter 6. https://www.boi.org.il/en/NewsAndPublications/RegularPublications/Pages/DochBankIsrael2018.aspx.

Baron, E. (2022). School spending and student outcomes: Evidence from revenue limit elections in wisconsin. *American Economic Journal: Economic Policy 14*(1), 1–39.

Barrios-Fernández, A. and G. Bovini (2021). It's time to learn: School institutions and returns to instruction time. *Economics of Education Review 80*, 102068.

Biasi, B. (2021). The labor market for teachers under different pay schemes. *American Economic Journal: Economic Policy 13*(3), 63–102.

Biasi, B. (2023). School finance equalization increases intergenerational mobility. *Journal of Labor Economics 41*(1), 1–38.

Biasi, B., J. M. Lafortune, and D. Schönholzer (2024). What works and for whom? effectiveness and efficiency of school capital investments across the us. Technical report, National Bureau of Economic Research.

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. *Unpublished working paper, version May 19*, 2021.

Callaway, B. and P. H. Sant'Anna (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Chabrier, J., S. Cohodes, and P. Oreopoulos (2016). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *Journal of Economic Perspectives 30*(3), 57–84.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? Evidence from project STAR. *The Quarterly Journal of Economics 126*(4), 1593–1660.

Cohen, S. (2011). *Teachers' responsiveness to change: Factors affecting teachers reform Join 'New Horizon' in Israel.* The Hebrew University Jerusalem.

Cohodes, S. R. and K. S. Parham (2021). Charter schools' effectiveness, mechanisms, and competitive influence.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–2996.

De Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers (2018). Double for nothing? Experimental evidence on an unconditional teacher salary increase in Indonesia. *The Quarterly Journal of Economics 133*(2), 993–1039.

Dee, T. S. and B. Jacob (2011). The impact of no child left behind on student achievement. *Journal of Policy Analysis and management 30*(3), 418–446.

Fryer, J. R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *The Quarterly Journal of Economics 129*(3), 1355–1407.

Gilraine, M. (2020). A method for disentangling multiple treatments from a regression discontinuity design. *Journal of Labor Economics 38*(4), 1267–1311.

Gleason, P., M. Clark, C. C. Tuttle, and E. Dwoyer (2010). *The Evaluation of Charter School Impacts: Final Report. NCEE 2010-4029.* National Center for Education Evaluation and Regional Assistance.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Harris, D. N. and T. R. Sass (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics 95*(7-8), 798–812.

Hemmings, P. (2010). Israeli education policy: How to move ahead in reform. *Economics Department WP, OECD* (781).

Hendricks, M. D. (2015). Towards an optimal teacher salary schedule: Designing base salary to attract and retain effective teachers. *Economics of Education Review 47*, 143–167.

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

Hyman, J. (2017). Does money matter in the long run? Effects of school spending on educational attainment. *American Economic Journal: Economic Policy 9*(4), 256–280.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Jackson, C. K. (2020). *Does school spending matter? The new literature on an old question.* American Psychological Association.

Jackson, C. K., R. C. Johnson, and C. Persico (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *Quarterly Journal of Economics 131*(1), 157–218.

Jackson, C. K. and C. Mackevicius (2021). The distribution of school spending impacts. Technical report, National Bureau of Economic Research.

Jackson, C. K., W. C. . X. H. (2021). Do school spending cuts matter? evidence from the great recession. *American Economic Journal: Economic Policy 13*(2), 304–335.

Jepsen, C. and S. Rivkin (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources 44*(1), 223–250.

Johnson, R. C. and C. K. Jackson (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy 11*(4), 310–349.

Justman, M. (2018). Randomized controlled trials informing public policy: Lessons from project STAR and class size reduction. *European Journal of Political Economy 54*, 167–174.

Kedagni, D., K. Krishna, R. Megalokonomou, and Y. Zhao (2021). Does class size matter? How, and at what cost? *European Economic Review 133*, 103664.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental analysis of neighborhood effects. *Econometrica 75*(1), 83–119.

Lafortune, J., J. Rothstein, and D. W. Schanzenbach (2018a). Educational resources and student achievement: Evidence from the save harmless provision in new york state. *Economics of Education Review 66*, 167–182.

Lafortune, J., J. Rothstein, and D. W. Schanzenbach (2018b). School finance reform

and the distribution of student achievement. *American Economic Journal: Applied Economics 10*(2), 1–26.

Lavy, V. (2020). Expanding school resources and increasing time on task: Effects on students' academic and noncognitive outcomes. *Journal of the European Economic Association 18*(1), 232–265.

Leuven, E. and S. A. Løkken (2020). Long-term impacts of class size in compulsory school. *Journal of Human Resources 55*(1), 309–348.

Maagan, D. (2015). Changes in the wages of teachers following the New Horizon reform: Did the reform improve the salaries of teachers and the quality of applicants for teaching? Technical report, The Central Bureau of Statistics Israel Presentation [in Hebrew].

OECD (2009). Education at a glance 2009. OECD Indicators, OECD Publishing, Paris. https://www.oecd-ilibrary.org/education/education-at-a-glance-2009_eag-2009-en.

OECD (2015). Education at a glance 2015. OECD Indicators, OECD Publishing, Paris. https://www.oecd-ilibrary.org/education/education-at-a-glance-2015_eag-2015-en.

RAMA (2008). Evaluation of the New Horizon reform in the 2007/2008 school year. Technical report, National Authority for Measurement and Evaluation in Education Publishing [RAMA], Tel Aviv [in Hebrew].

RAMA (2012a). Evaluation of the New Horizon reform after four years of implementation: Qualitative research among school teams in the 2011 school year. Technical report, National Authority for Measurement and Evaluation in Education Publishing [RAMA], Tel Aviv [in Hebrew].

RAMA (2012b). Evaluation of the New Horizon reform in elementary and junior high education after three years of implementation. Technical report, National Authority for Measurement and Evaluation in Education Publishing [RAMA], Tel Aviv [in Hebrew].

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

Schueler, B. E., C. A. Asher, K. E. Larned, S. Mehrotra, and C. Pollard (2022). Improving low-performing schools: A meta-analysis of impact evaluation studies. *American Educational Research Journal 59*(5), 975–1010.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

ONLINE APPENDIX

Appendix A Data Description

Our data come from the 2005–2012 Growth and Effectiveness Measures for Schools (GEMS) program. GEMS is administered by Israel's National Authority for Measurement and Evaluation in Education and involves both tests in core subjects and questionnaires distributed to students, teachers and principals. The GEMS testing program is used to assess school progress. Individual GEMS scores are not released to students or school administrators.

GEMS tests and questionnaires are usually administered sometime between mid-March and mid-June (except for the 2005 and 2006 school years, when the tests were given in October or November). The GEMS are drawn from a representative 1-in-2 sample of all elementary and middle schools in Israel, so that each school participates in the GEMS once every two years. In this study we use only elementary school data since adoption rates in lower-secondary schools were very low in the first years of the reform. In addition, since 2011 another reform has been implemented in secondary schools, complicating any attempt to analyze the effect of the NH reform in lower-secondary schools.

A.1 The GEMS Tests Sample

The GEMS tests in math, science, native language skills (Hebrew or Arabic reading and writing), and English are given to fifth-graders (elementary school) and eighth-graders (middle school). In principle, all students except those in special education classes are tested; in practice, the proportion of students tested is above 90 percent. In 2005 and 2006, participating schools were tested in four subjects. Since 2007, only two subjects at a time are tested, either math and language (Hebrew/Arabic) or science and English.

Our eight-year sample (2005–2012) includes test scores of 196,550 fifth-graders, or about 115,000 students for each subject from 877 schools. In our sample, the average exams attrition rate is 8 percent. Estimates are similar when the sample is limited to classes in which at least 50 percent of students were tested. Attrition is unrelated to the NH reform.

A.2 The GEMS Student Questionnaires Sample

The GEMS student questionnaires are distributed to all fifth- to ninth-graders in GEMS-participating schools. Since the current study focuses on elementary schools, we use only fifth- and sixth-graders' data.[30] The student questionnaires are anonymous and deal with various aspects of the school and the learning environment. Students are asked to rate the extent to which they agree with a series of statements on a five-point scale ranging from *strongly disagree* (1) to *strongly agree* (5). Statements related to the frequency of involvement in violent events (bullying) in the last month were rated on a three-point scale: *three times or more* (1), *once or twice* (2), and *never* (3). These statements are recoded into dummy variables indicating whether the student suffered from bullying at least once in the last month.

We use twenty-eight statements from the GEMS student questionnaires to measure school-learning environment outcomes.[31] We group the statements into six main indices: personal relations between students and teachers; teachers' efforts to help students; general satisfaction with the school; students' misbehavior in class; suffering from physical bullying; and suffering from social bullying. See Table A1 for a list of the statements by the six indicators. We construct the first four indices using a similar method as used by Kling et al. (2007). We standardize a class's mean answer to each statement by year and grade and then take a simple (within-class) mean of the resulting standardized variables to construct the index. The bullying indices were built by using a simple (within-class) mean of the class's mean answer to each statement (indicating the percentage of students in the class who suffered from bullying at least once in the last month).

[30] Elementary school in Israel runs from first to sixth grade.

[31] The full GEMS student questionnaires include more than 80 questions. We use only questions that were formulated identically in all years of the study.

Table A1: Statements Included in the School Learning Environment Indicators

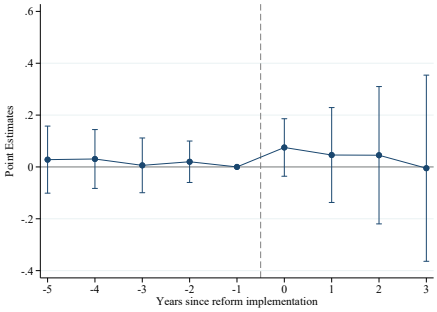
The indicators	The statements
Student–teacher relations	I have close ties with most of my teachers
	When I'm sad or feel bad, I feel comfortable talking about it with one of my teachers
	Most of my teachers care what happens to me, not only in connection with studies
Teachers' expectations of success	Most of my teachers give me the feeling I can do well in school
	Most teachers expect all students to make an effort in school
	Most teachers expect each student to improve his/her academic achievement
Teachers' efforts to help students	Most teachers explain to each student personally exactly what to do to improve his/her schoolwork
	When students have difficulty understanding the material, most teachers explain to them what they can do to better understand
	When teachers return our work or tests, most of them note what is correct and what needs to be improved in the pupil's answers
	When a student fails a class test or classwork, most of the teachers help them understand why it happened
Involvement in violent incidents	Last month I was hit, kicked or punched by a student who wanted to hurt me
	Last month a student hit me hard
	Last month a student pushed me
	Last month a student used a stick, a stone, a chair or another object to hurt me
	Last month a student threatened to hurt me at school or after school
	Last month a student tried to persuade other students not to talk to me
	Last month a student spread false rumors about me to hurt me
	Last month I was excluded. A group of students did not want to talk or play with me
Student misbehavior in class	Often students make noise, mess around in the classroom and interfere with learning
	Students in my class answer back to teachers
General school satisfaction	I like being at school
	If I could, I would go to another school (reverse-coded)

Our six-year sample (2007–2012) includes 11,005 fifth- and sixth-grade cohorts from 877 schools. This sample covers an annual average of 430 Jewish public (secular and religious) schools and 1834 classes per year. The average attrition rate is less than 1 percent, and estimates are similar when (a) including only items that were completed by at least 70 percent of the students in each class, and (b) excluding students who did not complete three or more items out of the twenty-eight.

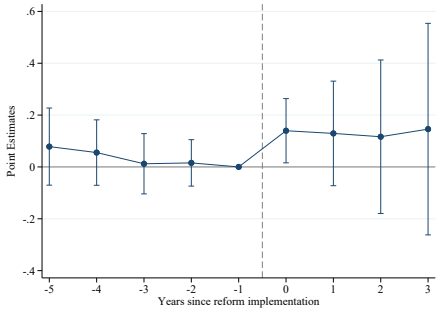
Appendix B Robustness Checks and Additional Results

Figure B1: Effect of the NH Reform on Students' and Schools' Pre-determined Characteristics

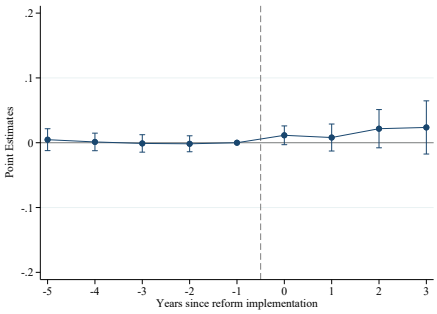
(a) Father's Years of Education



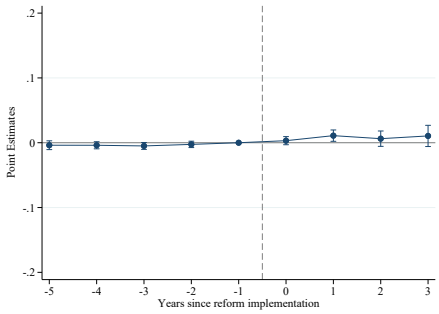
(b) Mother's Years of Education



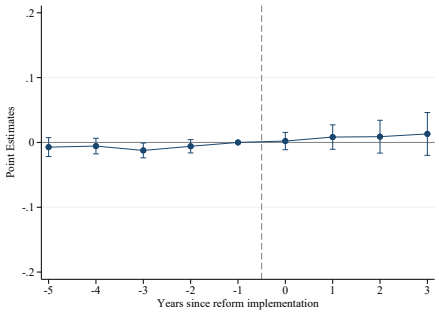
(c) Number of Siblings



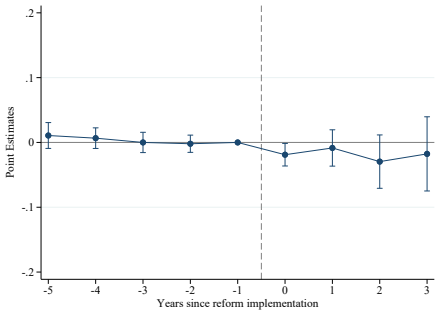
(d) Israeli Ethnicity



(e) Class Enrollment



(f) Grade Enrollment

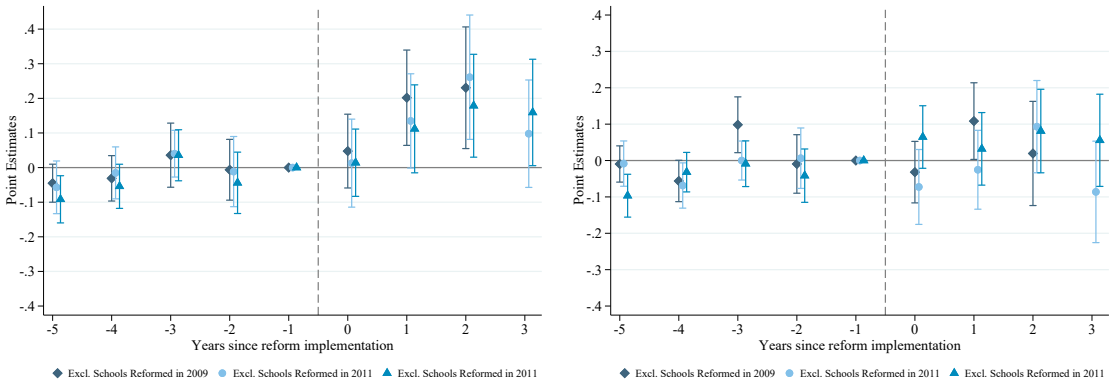


Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variables are students' and schools' time-varying predetermined characteristics: the student's father's years of schooling (panel a), the student's mother's years of schooling (panel b), the student's number of siblings (panel c), indicator for Israeli ethnicity (panel d), log of school's class size (panel e) and log of school's grade enrollment (panel e). Regressions are estimated using OLS and include the full set of event time dummies, school fixed effects and year fixed effects. The sample includes fifth-grade students from 877 Jewish (Hebrew-speaking) state elementary schools that participated in the GEMS tests between 2005 and 2012. Standard errors are clustered at the school level.

Figure B2: Effect of the NH Reform on Test Scores - Excluding School Cohorts by Year of Implementation

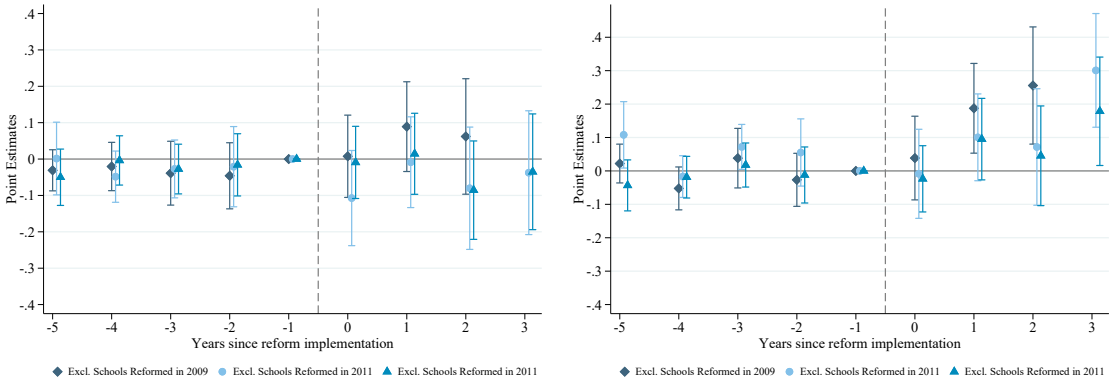
(a) Math Test Scores

(b) Language Test Scores



(c) Science Test Scores

(d) English Test Scores



Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variable is the student's standardized test score by year in math (panel a), language (b), science (c), and English (d). The dark-blue diamonds indicate the coefficients from a sample that excludes schools which reformed in 2009, the light-blue circles indicate the coefficients from a sample that excludes schools which reformed in 2010, and the blue triangles indicate the coefficients from a sample that excludes schools which reformed in 2011. All specifications estimate equation (1) using TWFE and include the full set of event-time dummies, school fixed effects, year fixed effects, district by year fixed effects, and controls for student and time-varying school characteristics. Student characteristics include a gender dummy, both parents' years of schooling, number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. Time-varying school characteristics include SES index, interaction between the SES index and a dummy for the post-reform period, and log of enrollment. All samples includes fifth-grade students from Jewish (Hebrew-speaking) state elementary schools that participated in the GEMS tests between 2005 and 2012. The dark-blue diamonds indicate the coefficients from a sample of 582 schools, the light-blue circles indicate the coefficients from a sample of 573 schools, and the blue triangles indicate the coefficients from a sample of 672 schools. Standard errors are clustered at the school level.