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

Yaniv Reingewertz and Adi Shany*

May 8, 2022

Abstract

This study analyzes the effect of a broad-based 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 on the school learning environment. We show that over the three years following introduction of the reform, test scores rose in math and English, and improvements were seen in various aspects of the school learning environment. The main mechanisms for the findings are suggested to be more small-group learning, better working conditions for some subject teachers, and an increase in the number of hours that teachers spend in school

Keywords: School reform, student achievements, school learning environment

JEL Codes: 121, 124, 128

^{*}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

Can school reforms effectively improve student outcomes? Under what conditions do these reforms succeed, and what inputs are required? These are among the most pressing questions in education economics. The literature offers some evidence on how specific inputs affect student achievement and the association between education reforms and education outcomes, but does not fully reveal the mechanisms by which school reforms influence the education process. This paper takes a step toward filling this gap.

This paper evaluates the effect of a broad-based national reform introduced in Israeli elementary and lower-secondary schools in 2008. This multi-faceted educational intervention, known as the New Horizon (NH) reform, aimed primarily to improve teaching quality. This was done partly by boosting teachers' status and raising their salaries in order to make the profession more attractive to talented newcomers and encourage greater engagement among existing teachers.¹ At the same time the reform aimed to address the problem of large gaps in student achievement and improve student-teacher relationships. Under the NH reform, teachers' working hours were lengthened to accommodate smallgroup teaching in exchange for more generous compensation. The reform also introduced new criteria governing teachers' professional development and promotions, as well as instituting a standard system for evaluating educational staff. We evaluate the short-term effect of the reform on student achievement and the school learning environment using a staggered difference-in-differences approach. We find that the reform improved student test scores in English and math over the four years from its implementation. There is also evidence to suggest that test scores in Hebrew (pupils' native language) improved among students from low socio-economic backgrounds. The reform also improved measures of the school learning environment such as student-teacher relations, student assessments of teachers' efforts, and students' behavior in class.

¹Before the reform, in 2007, the average pay of teachers in Israel was equivalent to 62 percent of GDP per capita. In 2013, five years after the reform's initial implementation, the average was equivalent to 100 percent of GDP per capita. According to OECD Indicators, OECD Publishing, Paris (2009) and OECD Indicators, OECD Publishing, Paris (2015), 100 percent of GDP per capita falls toward the bottom end of the typical range for teachers' pay in most OECD countries.

Countries across the Western world are engaged in efforts to improve their schools, in response to social and economic expectations (e.g., to improve students' readiness to enter the world of work or further education, or to promote social mobility). As schools' most significant and costly resource, teachers are central to school improvement efforts, and specifically to efforts to better student achievement (Harris and Sass, 2011, Jackson et al., 2014, Hendricks, 2015, Koedel et al., 2015, Araujo et al., 2016, Barrios-Fernández and Bovini, 2021, Biasi, 2021 and Hemelt et al., 2021).² Accordingly, many researchers working in the economics of education would argue that teacher quality is the single most important educational factor that can be influenced by policymakers. Empirical evidence supports this, suggesting that teacher quality is possibly the most important school-related attribute explaining student achievement (Rivkin et al., 2005). Another important school input thought to affect student outcomes is class size, though its effect is contested in the literature. ³ Notably, though, the impact of any given change in the context of large-scale policy reforms may be difficult to evaluate due to general equilibrium effects.

Another strand of the literature evaluates more broadly the effects of large-scale, "whole-school" reforms, which seek to affect allocation of multiple educational resources simultaneously. This category includes school finance reforms, which aim to improve outcomes by increasing overall school spending. Recent quasi-experimental studies that relate school spending to student outcomes overwhelmingly support a causal relationship between the two.⁴ However, the question of what kinds of spending increases matter the most remains unresolved. The same is true for other types of whole-school reforms. For example, the Comprehensive School Reform (CSR) model implemented widely in the US over recent decades aimed at coordinating reforms to the curriculum, instruction, professional development, classroom management, school management, and parental involve-

²Teacher salaries usually constitute 50–60 percent of total spending by local school systems, making investment in teachers the single largest share of local school system budgets (Sibieta et al., 2018, Bayer et al., 2020).

³See, among others Angrist and Lavy (1999), Hoxby (2000), Jepsen and Rivkin (2009), Chetty et al. (2011), Chingos (2013), Justman (2018), Angrist et al. (2019), Leuven and Løkken (2020), Gilraine (2020), Kedagni et al. (2021).

⁴Jackson et al. (2016), Hyman (2017), Lafortune et al. (2018), Jackson (2018), Johnson and Jackson (2019), Jackson (2020) and Jackson and Mackevicius (2021).

ment. A number of innovative programs have been introduced under the CSR rubric, including Success for All and Direct Instruction (for an overview on CSR see Slavin, 2007). Yet while there is some evidence that such reforms improve academic outcomes (Borman et al., 2003, Walpole et al., 2017, Bonilla and Dee, 2020), it remains unclear which specific mechanisms are responsible for the effectiveness of the reform.

This paper presents an empirical investigation of how the New Horizon reform affected educational and school learning environment outcomes in the short term, using a sample which covers all Jewish non-ultraorthodox elementary schools in Israel over the years 2005–2012.⁵ Our outcome data are drawn from fifth-grade test scores on national standardized exams in math, 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. 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. To address recent concerns raised by the econometric literature on DiD models with staggered adoption (see, among others Sun and Abraham, 2020,

⁵While the reform applied to both elementary and lower-secondary schools, adoption rates among the latter were very low in the first years of the reform. Hence, in this study we use only elementary school data. Arab schools and Ultraorthodox schools are operating as separate streams and are thus not included in our sample.

Callaway and Sant'Anna, 2020, De Chaisemartin and d'Haultfoeuille, 2020, Goodman-Bacon, 2021 and Borusyak et al., 2021), we show that our results are robust to using an alternative estimator proposed by Sun and Abraham (2020).

The results suggest that over the study period, the NH reform had a positive and significant short-run effect on both math and English test scores and the school learning environment. As of the final year studied, 2012, math test scores had increased by about 0.13 standard deviations and English test scores by about 0.20 standard deviations. Overall, test scores in Hebrew language and science did not improve as a result of the NH reform. In terms of environmental outcomes, the reform improved the relationship between teachers and students, increased teachers' effort to help their students, and improved students' behavior in class. However, we do not find an effect on bullying, whether physical or social. 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 Hebrew language test scores (i.e., reading and writing). However, we do not find a different effect in English test scores between students with different level of parents' education. These results may be explained by the higher rates at which students from lower educational backgrounds are included in small-group learning, which is mainly devoted to math and Hebrew language. 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.

This paper offers two main contributions to the literature on school reform. First, the paper assesses the effect of a national "whole-school" reform on student achievement. Notably, while school finance reforms generally deliver additional resources to disadvantaged school districts in an attempt to close academic achievement gaps, and CSR is

implemented as different programs in different schools, the NH reform is a wide-ranging uniform reform encompassing nearly the entire elementary-school system in Israel. Thus, this study offers the opportunity to evaluate the effect of a whole-school reform which is implemented the same way in all traditional state-controlled schools across the whole socio-economic status distribution. Hence, the findings may have general implications that are relevant not only for disadvantaged schools. Moreover, the paper offers suggestive evidence regarding the mechanisms by which school reforms affect student outcomes. Many studies of educational reforms focus solely on outcomes without giving much thought to the mechanisms at play. Other studies examine specific inputs in an isolated manner, as opposed to in the context of a broader reform.⁶

Second, and relatedly, this paper also evaluates the effect of the reform on nonacademic outcomes like the school learning environment. Such 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, 2018). To the best of our knowledge, we are the first to evaluate the effect of school reform 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 New Horizon educational reform agreement was signed in March 2008 between the Israeli Ministry of Education and the teacher's union, which represented all state ele-

⁶for example, teachers training (Angrist and Lavy, 2001), length of the school year (Pischke, 2007), teacher incentives and pupil-teacher ratios (Duflo et al., 2015), teachers salary (De Ree et al., 2018) and instructional time **?**.

⁷Gleason et al. (2010), in their study of charter schools, also evaluate non-academic outcomes such as student effort in school, behavior, and attitudes, as well as parental involvement and satisfaction.

mentary and some lower-secondary school teachers. All target schools were fully funded by the state, encompassing Hebrew-speaking secular state schools, state religious schools, and Arabic-speaking state schools.⁸ The reform was implemented gradually, with roughly twenty percent of the elementary schools enrolling each year from 2008 to 2012. The reform had five main objectives: to (a) improve teacher quality by boosting teachers' status and raising their salaries; (b) reduce gaps in education outcomes by providing opportunities for disadvantaged students; (c) raise overall student achievement; (d) improve the learning environment; and (e) reinforce and broaden the scope of the school principal's authority (RAMA, 2012a, RAMA, 2012b).

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).⁹ In return for the increased workload, teachers received a wage increase averaging twenty-six percent (Cohen, 2011). Yet as the wage increase was accompanied by increased time spent in school, teachers' hourly wage did not increase by the same margin.

The reform also introduced new criteria governing teachers' professional development and promotions, as well as instituting a standard system for evaluating all educational staff (teachers, vice-principals, and principals). Additional pay grades were added to allow teachers more opportunities for promotion in return for professional development. In addition, 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 also mandated improvements in teachers' physical environment, in particular the provi-

⁸Israel's education system reflects the state's multicultural population. Most children attend secular state schools, with instruction in Hebrew. State religious schools include extra teaching in Jewish studies and observance. Together, these groups comprise the mainstream Jewish state school system examined in the present study. Arab and Druze state schools teach in Arabic and focus on Arab and Druze history, religion, and culture. A fourth group of students attend private schools, many serving the country's ultra-orthodox community.

⁹The 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.

sion of suitable work areas. With respect to principals, the reform introduced a new and substantially more generous pay scale, and expanded principals' scope of responsibilities, giving them greater control over which teachers are hired, promoted, tenured, and fired.

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.¹⁰Beginning with the reform's second 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.¹¹ At this stage, once a school joined 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.

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. 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. The GEMS data that we use include fifth-graders' test scores in reading and writing (i.e., Hebrew), mathematics, science, and English, along with questionnaires polling fifth- and sixth-graders on their school learning environment, and questionnaires distributed to all teach-

¹⁰Forty percent of these schools had already undergone a previous school reform, and many strongly supported the educational vision embodied by the NH reform (RAMA, 2008).Teachers who wished to continue working under the same terms as before the reform in these schools were free to do so for the first two years of the reform (Cohen, 2011).

¹¹The Israeli education system is divided into seven districts, with district managers responsible for implementing the reform in their district.

ers and principals. The available students' questionnaires date back to 2007, while those distributed to teachers and principals date only to 2009.¹² 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).¹³ 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 statistical analysis looks at fifth- and sixth-grade pupils in the Jewish state school system, including both secular and religious schools.¹⁴ The analysis excludes schools that joined 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).¹⁵ We also exclude 12 schools which 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

¹²Students' 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.

¹³The 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).

¹⁴As noted earlier, the sample is restricted to Jewish state schools that follow the same national curriculum and participate in the GEMS national testing. For these reasons we exclude Arab schools and private Jewish schools.

¹⁵As mentioned in section 2, schools that implemented the reform in 2008 differed from schools that implemented the reform afterword. We exclude schools that took part in a previous reform because it would not be possible to separate the effects of the two reforms.

scores for 196,549 fifth-grade students from 877 schools. About 115,000 students were tested in math and Hebrew 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 onward, schools that tested students in math and Hebrew in a given year were not the same schools which tested in science and English in the same year. Hence, not all students tested in all subjects.

To protect students' privacy, the fifth- and sixth-grade GEMS questionnaires are not identified by student. Hence, we aggregated the data to the class level by calculating an average value of the students' 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 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 a detailed description of the GEMS student questionnaires and 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 fifthgrade test scores 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 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

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. (2018) educational reforms may affect students' outcomes - in particular academic achievement outcomes - slowly and gradually rather than immediately, 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 difference-in-differences estimator, as well as a parametric DiD proposed by Lafortune et al. (2018) 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, Hebrew, science and English by estimating the following event study specification separately for each subject:

$$y_{ist} = \sum_{j \neq -1} \beta_j \mathbb{1}[t = t_s^* + j] + \alpha_s + \delta_t + \lambda S_{st} + \gamma X_{ist} + \rho_{st} + u_{ist}$$
(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 are 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.¹⁶ 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 school. 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-2008,¹⁷ and log enrollment; and X_{ist} , which is a vector of characteristics of

¹⁶As 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 prior to treatment.

¹⁷The SES index formula was changed in 2008.

student *i* in school *s* in year *t*. Student characteristics include a gender dummy, both parents' years of schooling, 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} .¹⁸

Since the school learning environment dataset is available only from 2007 and we observe each school only once in the pre-reform years, we cannot perform the event study analysis for this sample. Hence, we estimate a standard difference-in-differences specification for both test scores outcomes and learning environment outcomes as the following:

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

where *i* denotes the unit of analysis which, depending on the outcome, may be 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_i st 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 implementation of the reform. In implementing a staggered difference-in-differences analysis we follow the analytic approach used by Lafortune et al. (2018) to accommodate both an immediate and a gradual effect. This approach is presented in Equation 3:

$$y_{ist} = \beta_1 Reform_{st} + \beta_2 Reform_{st} * Reform_trend_{st} + \alpha_s + \delta_t + \lambda S_{st} + \gamma X_{ist} + \rho_{st} + u_{ist}$$
(3)

where $Reform_trend_{st}$ is a trend variable which measures the years since the reform was implemented in a given school; it can take negative values before the implementation.¹⁹

¹⁸In Section 5 we also show results from specifications that do not include district by year fixed effects or include district linear time trend instead.

¹⁹Lafortune et al. (2018) also include $Reform_trend_{st}$ as a control, and suggest it represents a falsifica-

Here β_2 captures delayed effects of the reform, and represents the annual change in outcomes in school *s* from the implementation year, relative to the same school prior to the reform. Throughout our analyses, we use standard errors clustered at the school level.

To interpret the results of the staggered difference-in-differences specification as a causal treatment effect of the NH reform on student outcomes, we rely on two key identifying assumptions: no anticipation of the treatment, and parallel trends (Sun and Abraham, 2020, 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 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.²⁰ 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 which would change their work environment in year t + 1. In any case, should such anticipation effects exist, they will attenuate our estimates toward zero. As for the students, we see no reason to believe that elementary school students were aware of the reform before it came into effect.

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. 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 these coefficients do not differ significantly from zero.

tion 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.

²⁰Nor could students move between schools to avoid the reform, thus we do not expect anticipatory effects of parents.

Recent econometric studies have highlighted that event study coefficients may be biased 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, 2020). In these cases, each event time coefficient may be "contaminated" with effects from other cohorts. Specifically, Sun and Abraham, 2020 show 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).²¹ 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 test the robustness of our results to using an alternative estimator proposed by Sun and Abraham (2020) that allows for heterogeneous treatment effects. We show that our results are consistent when using this alternative estimator, further suggesting 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, Hebrew, science, and English (panels a, b, c and d respectively). The solid lines

²¹As shown by Goodman-Bacon (2021) the inclusion of a control group can partially help to alleviate this issue. Unfortunately, as almost all schools reformed by 2012, we are unable to include never-treated schools as a control group.

represent estimates from the event study specification (equation 1), while the dashed lines show pointwise 90 percent confidence intervals. There are no indications of differential trends in test scores in any subject prior to the reform implementation. We also do not see an immediate reaction in the first year of the reform's implementation (t = 0), but math and English test scores begin an upward trend starting 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). The effect on English test scores takes more time to materialize, but in the fourth year there is a large increase in test scores of about 0.2 standard deviations (t = 3). In no case is there any sign of a pre-event trend that would suggest a violation of our quasi-random timing assumption.

To provide additional supporting evidence for the validity of our event study estimates, we repeated the analysis using the interaction-weighted estimator proposed by Sun and Abraham (2020), which is robust to heterogeneous treatment effects. The interaction-weighted estimator is a regression-based estimator that provides a weighted average of the treatment effects, producing estimates that are easier to interpret than those from a standard two-way fixed effects estimator (Sun and Abraham, 2020). Specifically, each event time coefficient from this estimation is a weighted average of the cohort-specific average treatment effect on the treated, where the weights are given by the share of cohorts that experienced at least t periods relative to the treatment, and normalized by the total event time periods estimated, where the schools who joined the reform in the last year (2012) serve as a control group. The results are presented in Figure B1 in the Online Appendix. As can be seen, the results using Sun and Abraham (2020) methodology are qualitatively similar to the two-way fixed effects event study results (Figure 1).

Table 3 reports the estimates of equations (2) and (3) for the standardized test scores in the four GEMS subjects (math, Hebrew, 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.

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 scores in Hebrew and science. Moreover, the effects in math and English are gradual rather than immediate. The estimates are similar across all specifications. According 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 why the NH reform affects students' performance in this manner, and why the effect is seen only in math and English and not in Hebrew and science. The results presented in Table 3 are very similar to those of Figure 1, and essentially give an average post-reform effect for the year effects presented in Figure 1.

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 teachers' 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 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 pre-reform trend in test scores and in teachers characteristics, as shown in Section 6.2, provides confidence 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. Given its positive effect on math and English test scores and the school learning environment, it is important to understand which of these inputs may be driving our results. In this section we discuss the potential contributions of the main elements of the reform. 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.

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 reinforcement of core subjects, especially reading and writing in the students' 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 bellow to 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 Hebrew test scores due to the reform. Math and Hebrew 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 Hebrew, the students who came from homes characterized by lower education levels, and who were therefore likely themselves 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 Hebrew and math than the other subjects, the findings provide suggestive evidence that the effect on math and Hebrew 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 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 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

terms, 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 potential mechanisms 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 school. Figure 2 reports the results for the two-way fixed effects estimation and Online Appendix Figure B2 reports the results using Sun and Abraham (2020) estimation. The two-way fixed effects 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 B2 suggest that principals with a slightly higher tenure (of 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.

Figure 3 and Online Appendix Figure B3 present the estimates for the teachers' composition in school for the two-way fixed effects estimation and the Sun and Abraham (2020) estimation, respectively. 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) are the percentage of new teachers (less than 5 years) and veteran teachers (over 30 years) in school, respectively. The results show that teachers' tenure moderately increased after the reform, suggesting that turnover rates fell, although the coefficients are marginally statistically significant. In Figure B3 The estimates for teachers' tenure are not statistically different from zero. The reform seems to have had an effect on the academic qualifications of teachers, as the percentage of teachers with an academic degree significantly increased after the reform. Figure B3 shows a large increase of 20 percentage points in the share of teachers with an academic degree while Figure 3 shows a modest increase of only 4 percentage points after three years of the reform. This 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. The post-reform changes in the percentages of new teachers and veteran teachers are not precise in Figure 3, but we do observe an increasing trend in the proportion of veteran teachers and a decline in the proportion of new teachers. Figure B3 also shows a significant drop in the percentage of new teachers in the first year after the reform was implemented. For all these variables, we observe no pre-event trend both in Figure 3 and Figure B3. This strengthens our assumption that the changes seen in our measured outcomes following the reform are not linked to anticipation effects or other pre-reform events.

6.2.2 Differences by Teaching Subject

Our main analysis indicates that the reform contributed to improvements in students' achievement in math and English. We provided suggestive evidence that the effect of the reform on math test scores is due to the increase in small-group learning. Still, small-group learning does not explain the improvement in English test scores. One possible explanation, which will be assessed here, is that the improvement in English test scores might derive from effects of the reform on teachers. In order to further explore these

effects, we use the teachers' GEMS questionnaires to provide some information on the differences between teachers in different subjects.

Teachers in schools that participated in 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, Hebrew, science, and English), and focus on teachers who teach these subjects to children 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, Hebrew, 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 pre- and post- reform for the subset of schools that we observe in the teachers' GEMS questionnaires.

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 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 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.

6.2.3 Teachers Working Hours

An important component of the reform was the increase in teachers' paid working hours. The administrative data in our dataset allow us to observe working hours for each teacher in each school. An interesting test is whether the change in working hours was similar for all teachers, as differences between teachers in different subject areas may offer a possible explanation for our results on the test scores. Figure 4 shows the results for estimating equation (1) where the dependent variable in panel A is average working hours 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.²²

The NH reform considerably increased the working hours of teachers, by almost 8 weekly hours on average (panel A). However, this marked increase in working hours conceals differences between teachers of different subjects. Specifically, if we compare homeroom teachers with English teachers, we see that the increase in working hours experienced by English teachers was lower than that of homeroom teachers by roughly two hours (panel B). These results support our conclusion (from Table 7) that English teachers may have benefitted the most from the reform, in terms of gaining in professional recog-

 $^{^{22}}$ The increase in the working hours of teachers is a mechanical results of the reform - one of the attributes of the reform is the requirement that teachers would increase their working hours. Therefore, we do not apply a correction for heterogeneous treatment effect by using Sun and Abraham (2020) estimation.

nition and pay without the increased stress likely to accompany greater responsibilities at school. This again provides a potential channel for the influence of the reform on 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 teachers, who usually teach math and Hebrew, 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 and Hebrew 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, which could be another mechanism for a uniform positive effect on general school environment.

7 Conclusions

New Horizon was a nationwide educational reform that was gradually implemented in Israeli elementary schools starting in 2008. Its main aim was to improve academic achievement and narrow academic disparities between lower- and higher-performing students. This study evaluated the short-term effects of the reform. The results, based on fifthgrade students' test scores, clearly indicate that New Horizon had a positive and significant short-term effect on student performance in core subjects, mainly English and math. The evidence also points to a heterogeneous effect, as students whose parents have low levels of education (a proxy for student ability) show higher gains in math and Hebrew than students in their classes whose parents are better educated. 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).²³ Additionally, the reform helped to improve the school learning environment, as reflected in student questionnaires among fifth- and sixth-graders. The evidence suggest that New Horizon improves student–teacher relations, improves teachers' efforts to help students, and reduces misbehavior in class (as viewed by students). During the study period, the reform did not reduce student involvement in violent incidents (bullying) and had no effect on students' overall satisfaction with their schools.

This is the first paper that provides a detailed evaluation of the causal effects of the New Horizon 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 suggest that the positive effect of the reform on students' academic achievement probably traces mainly 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 may be greater than our estimates, which relate to the overall effect of the reform.

²³Back-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.

The implications of this study are relevant for policymakers not only in Israel, but may offer guidance for the implementation of educational reforms in other countries. The New Horizon reform focused mainly on teachers' working conditions and added individual instruction time. We therefore explored the effects of these specific components within the school reform context, something the literature has not previously been able to isolate. In addition, this is one of the few studies to analyze a nationwide reform, with the result that its implications are more likely to have external validity.

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, New Horizon may create a system that attracts and retains more able teachers and produces larger positive effects in the long run. However, future research on the long-term effects of the reform and an in-depth cost–effectiveness analysis is 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.
- 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.
- Bayer, P. J., P. Q. Blair, and K. Whaley (2020). Are we spending enough on teachers in the US? *NBER Working Paper* (w28255).
- Biasi, B. (2021). The labor market for teachers under different pay schemes. American Economic Journal: Economic Policy 13(3), 63–102.
- Bonilla, S. and T. S. Dee (2020). The effects of school reform under NCLB waivers: Evidence from focus schools in Kentucky. *Education Finance and Policy* 15(1), 75– 103.

- Borman, G. D., G. M. Hewes, L. T. Overman, and S. Brown (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research* 73(2), 125–230.
- 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*.
- 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.
- Chingos, M. M. (2013). Class size and student outcomes: Research and policy implications. *Journal of Policy Analysis and Management*, 411–438.
- Cohen, S. (2011). *Teachers' responsiveness to change: Factors affecting teachers reform Join 'New Horizon' in Israel.* The Hebrew University Jerusalem.
- 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.
- Duflo, E., P. Dupas, and M. Kremer (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics 123*, 92–110.
- 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.
- Hemelt, S. W., H. F. Ladd, and C. R. Clifton (2021). Do teacher assistants improve student outcomes? Evidence from school funding cutbacks in North Carolina. *Educational Evaluation and Policy Analysis* 43(2), 280–304.
- 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., J. E. Rockoff, and D. O. Staiger (2014). Teacher effects and teacherrelated policies. *Annual Review of Economics* 6(1), 801–825.
- 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.
- Koedel, C., K. Mihaly, and J. E. Rockoff (2015). Value-added modeling: A review. *Economics of Education Review* 47, 180–195.
- Lafortune, J., J. Rothstein, and D. W. Schanzenbach (2018). 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 Indicators, OECD Publishing, Paris (2009). Oecd education at a glance 2009. https://www.oecd-ilibrary.org/education/ education-at-a-glance-2009_eag-2009-en.
- OECD Indicators, OECD Publishing, Paris (2015). Oecd education at a glance 2015. https://www.oecd-ilibrary.org/education/ education-at-a-glance-2015_eag-2015-en.
- Pischke, J.-S. (2007). The impact of length of the school year on student performance and earnings: Evidence from the German short school years. *The Economic Journal* 117(523), 1216–1242.
- RAMA (2008). Evaluation of the New Horizon reform in the 2007/2008 school year.Technical report, National Authority for Measurement and Evaluation in EducationPublishing [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., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica* 73(2), 417–458.
- Sibieta, L., C. Belfield, et al. (2018). 2018 annual report on education spending in england.
- Slavin, R. E. (2007). Comprehensive school reform. Thousand Oaks, CA: Sage.
- Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.
- Walpole, S., S. Amendum, A. Pasquarella, J. Z. Strong, and M. C. McKenna (2017). The promise of a literacy reform effort in the upper elementary grades. *The Elementary School Journal 118*(2), 257–280.



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

(a) Math Test Scores

(b) Hebrew 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), Hebrew (b), science (c), and English (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 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. 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 2: Effect of the NH Reform on Principals' Characteristics - 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 principals' average tenure (panel a) and a dummy variable for whether principals are new at their school (b). 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.





(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: 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.

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



(a) Average Working Hours (All Teachers)

(b) Difference in Working Hours between Homeroom Teachers and English Teachers

Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variable is 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 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.

	Hebrew-speaking	Hebrew-speaking	Arabic-speaking	
School year	state schools (Jewish	state schools (Jewish	state schools	Total
	secular)	religious)		
	(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

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

	Test S San	Scores nple	School I Environme	Learning ent Sample
	Mean	S.D.	Mean	S.D.
Pa	nel A: School C	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	87	17	87	17
Panel	B: Student/Cla	ss Characteristi	cs	
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	11,0	008
J	Panel C: Outcor	ne Variables		
I. Test Score Outcomes				
Math score [N=115.360]	65.79	21.60		
Hebrew 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 Ot	utcomes			
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= 11,00	8]		-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 0081		0 146	0.077

Table 2: Descriptive Statistics

		Math			Hebrew			Science			English	
	(1)	(2)	(3)	(4)	(5)	(9)	(7)	(8)	(6)	(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	0.20
Panel B:	(((0,0))	(((())))	(((())))	(7000)	(000.0)	(0000)	(0000)		(((())))	(0+0.0)	(740.0)	(0±0.0)
Post-reform	0.046 (0.039)	0.043 (0.039)	0.049 (0.039)	0.021 (0.033)	0.012 (0.034)	0.022 (0.034)	-0.012 (0.039)	-0.023 (0.040)	0.001 (0.040)	-0.008 (0.041)	-0.002 (0.043)	0.0005 (0.041)
Post-reform Trend	0.035** (0.018)	0.043** (0.020)	0.039** (0.019)	0.002 (0.016)	0.010 (0.017)	0.005 (0.016)	0.009 (0.022)	0.005 (0.021)	0.002 (0.021)	0.064^{***} (0.021)	0.067*** (0.021)	0.057*** (0.021)
Student Controls	X	X	X	×	×	×	×	×	×	×	×	X
School Controls	X	X	X	X	X	X	X	X	X	X	X	X
District by Year FE		X			Х			Х			Х	
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 estir	mates of par	ametric diff at score by y	erence-in-di	ifferences m	nodels corre 1–3) Hehre	sponding to	equation (2) scient) in panel A	and equations and	on (3) in pan H Enolish (c	iel B. The de	pendent
specifications include school	and year fiv	xed effects.	Student con	itrols includ	le a gender i	ndicator, bc	oth parents'	years of sch	ooling, nun	nber of sibli	ngs, a born-	in-Israel
indicator, and ethnic-origin	indicators.	School con	trols includ	e SES inde	x, interactic	on between	the SES inc	dex and a d	ummy for t	the post-ref	orm period,	and log
school enrollment. Standard * $p < 0.10$, ** $p < 0.05$, *** p	errors are c $p < 0.01$	lustered at i	the school le	evel.								

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

39

	Stude	ent-teacher rela	ttions	Teachers	' efforts to help	students	Genei	ral school satist	faction
	(1)	(2)	(3)	(4)	(5)	(9)	(7)	(8)	(6)
Panel A:	0.092***	0.092***	0.097***	0.111***	0.117***	0.117***	0.026	0.024	0.031
Post-reform	(0.033)	(0.033)	(0.033)	(0.034)	(0.034)	(0.034)	(0.038)	(0.038)	
Panel B:	0.103***	0.102***	0.108***	0.117***	0.122***	0.122***	0.040	0.037	0.044
Post-reform	(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 School Controls District by Year FE District 1 inser Time Trand	X X	X X X	X X X	X X	x x x	X X X	XX	XXX	X X X
Classes Schools		11,005 877	\$		11,005 877	<		11,005 877	\$
Notes: The table reports estin	nates of parame	etric difference-	-in-differences I	models correspondences of the series of the	onding to equat	ion (2) in panel	A and equation	(3) in panel B.	The dependent
variables are standardized sc.	ores for the cla	ss average in ea	ach of the schoc		ronment indica	tors. All specific	cations include	school and yea	rr fixed effects.
Student controls are class av	rerages of boys	b, both parents'	years of schoc		of siblings, a by	prn-in-Israel inc	licator, and eth	mic-origin indiv	cators. School
controls include SES index, i	interaction bety	veen the SES ir	ndex and a dum		t-reform period	, and log school	l enrollment. St	tandard errors a	ure clustered at

the school level. * p < 0.10, ** p < 0.05, *** p < 0.01

.

40

	Studen	t misbehavior i	n class	Suffered	from physical	bullying	Suffere	ed from social l	oullying
	(1)	(2)	(3)	(4)	(5)	(9)	(7)	(8)	(6)
Panel A: Post-reform	-0.104***	-0.103***	-0.110***	-0.004	-0.004	-0.004	-0.003	-0.004	-0.004
Panel B:	(0000)			(000.0)	(000.0)			(000.0)	(000.0)
Post-reform	-0.112*** (0.039)	-0.114^{***} (0.039)	-0.121*** (0.039)	-0.004 (0.003)	-0.004 (0.003)	-0.005 (0.003)	-0.003 (0.003)	-0.004 (0.003)	-0.004 (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	Х	Х
District by Year FE		Х			Х			Х	
District Linear Time Trend			X			X			x
Classes		11,005			11,005			11,005	
Schools		877			877			877	

á 5 2 2 the school level. * p < 0.10, ** p < 0.05, *** p < 0.01

.

41

		Math			Hebrew			Science			English	
	(1)	(2)	(3)	(4)	(5)	(9)	(2)	(8)	(6)	(10)	(11)	(12)
Post-reform	0.057 (0.039)	0.054 (0.038)	0.060 (0.039)	0.024 (0.034)	0.015 (0.034)	0.015 (0.034)	-0.013 (0.040)	-0.026 (0.040)	-0.001 (0.040)	-0.004 (0.042)	0.001 (0.044)	0.003 (0.042)
Post-reform* Low-educated Parents	-0.028 (0.023)	-0.032 (0.023)	-0.028 (0.023)	-0.004 (0.022)	-0.006 (0.022)	-0.006 (0.022)	0.010 (0.023)	0.010 (0.023)	0.014 (0.023)	-0.003 (0.021)	-0.004 (0.021)	-0.001 (0.021)
Post-reform Trend	0.021 (0.018)	0.027 (0.020)	0.024 (0.019)	-0.007 (0.017)	0.0004 (0.017)	0.0004 (0.017)	0.008 (0.021)	0.005 (0.021)	0.002 (0.021)	0.058^{***} (0.021)	0.061*** (0.021)	0.052** (0.021)
Post-reform Trend* Low-educated Parents	0.038^{***} (0.014)	0.040^{***} (0.014)	0.038^{***} (0.014)	0.025^{*} (0.013)	0.024^{*} (0.013)	0.024^{*} (0.013)	0.002 (0.015)	0.0008 (0.015)	0.00005 (0.015)	0.018 (0.013)	0.017 (0.013)	0.017 (0.013)
Student Controls School Controls District by Year FE District Linear Time Trend	x x	x x x	x x x	x x	x x x	x x x	X X	X X X	x x x	x x	x x x	x x x
Students Schools		115,360 877			114,860 877			114,382 872			112,928 870	
Notes: The table reports estifor students with parents' sc with the post-reform trend v (columns 7–9), and English	mates of pai hooling belc ariable. The (columns 1	rametric dif w the class dependent 0–12). All	ference-in-d ?s median v: variable is 1 specificatio	lifferences r alue, an inte the student' ns include a	nodels corre rraction of t s standardiz school and	esponding to his indicato ted test scor year fixed e	o equation () r with the po e by year ir ffects. Stud	3) with the ost-reform v n math (colu lent control	following ac /ariable, and umns 1–3), 1 s include a	lditional va l an interact Hebrew (co gender indi	riables: an i cion of this i lumns 4–6), lcator, both	ndicator ndicator science parents'

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

years of schooling, number of siblings, a born-in-Israel indicator, and ethnic-origin indicators. School controls include SES index, interaction between the SES index and a dummy for the post-reform period, and log school enrollment. Standard errors are clustered at the school level. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 7: Teacher Questionnaire Responses by Teaching Subject

	Math te	sachers	Hebrew	teachers	Science	teachers	English	teachers
	Pre-reform	Post-reform	Pre-reform	Post-reform	Pre-reform	Post-reform	Pre-reform	Post-reform
	(1)	(2)	(3)	(4)	(5)	(9)	(2)	(8)
Homeroom teacher $(1 = Yes, 0 = No)$	0.879 (0.326)	0.880 (0.325)	0.943 (0.231)	0.946 (0.227)	0.608 (0.488)	0.616 (0.486)	0.201 (0.401)	0.195 (0.396)
Main specialization in the subject $(1 = \text{Yes}, 0 = \text{No})$	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 $(1 = \text{Yes}, 0 = \text{No})$	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 $(1 = \text{Yes}, 0 = \text{No})$	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 $(1 = \text{Yes}, 0 = \text{No})$	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 $(1 = \text{Yes}, 0 = \text{No})$	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	3,184 414	2,429
Schools	414	414	414	414	414	414		414
Notes: The table reports means and standard deviat teaching subject and treatment status (pre- and post- schools that participated in the GEMS tests between 2	tions in parentl reform). The sc 2009 and 2012	heses for select ample includes at least twice, o	ed questions f teachers from	rom the GEMS Jewish (Hebrew e-reform period	teachers' que -speaking) stat and once in th	stionnaire by te elementary e post-reform		

period.

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 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 GEMSparticipating schools. Since the current study focuses on elementary schools, we use only fifth- and sixth-graders' data.²⁴ 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.²⁵ 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).

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

²⁴Elementary school in Israel runs from first to sixth grade.

²⁵The full GEMS student questionnaires include more than 80 questions. We use only questions that were formulated identically in all years of the study.

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.

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)

Appendix B Additional Results



Figure B1: Effect of the NH Reform on Test Scores - Sun and Abraham (2020) 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 in math (panel a), Hebrew (b), science (c), and English (d). Regressions are estimated using the method of Sun and Abraham (2020) 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. 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.





Notes: The figure plots the event time coefficients and their 90 percent confidence intervals from estimating equation (1) where the dependent variable is the principal's tenure (panel a) and a dummy variable for whether principals are new at their school (b). Regressions are estimated using the method of Sun and Abraham (2020) 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 ethnic-origin percentages. 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.

Figure B3: Effect of the NH Reform on Teachers' Characteristics - Sun and Abraham (2020) Estimation

(a) Teachers' Average Tenure



(b) Percentage of Teachers with Academic Degree



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 teacher's tenure (panel a), the percentage of teachers with an academic degree (b), the percentage of new teachers, defined as less than 5 years of experience (c), and the percentage of veteran teachers, defined as more than 30 years of experience (d). Regressions are estimated using the method of Sun and Abraham (2020) 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 ethnic-origin percentages. 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.