Economics of Education Review 94 (2023) 102383

Contents lists available at ScienceDirect

Economics of Education Review

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

Closing the income-achievement gap? Experimental evidence from high-dosage tutoring in Dutch primary education☆

Joppe de Ree [a,*], Mario A. Maggioni [b], Bowen Paulle [c], Domenico Rossignoli [b], Nienke Ruijs [d], Dawid Walentek [e]

[a] Erasmus University Rotterdam, The Netherlands
[b] DISEIS & CSCC Università Cattolica del Sacro Cuore, Milano, Italy
[c] University of Amsterdam, The Netherlands
[d] Dutch Inspectorate of Education and VU University Amsterdam, The Netherlands
[e] University of Warsaw, Poland

ARTICLE INFO

JEL classification:
H0
I21
I24

Keywords:
Education policy
High-dosage tutoring
Inequality of opportunity
Primary education
Randomized control trial

ABSTRACT

We present experimental evidence on a high-dosage math tutoring (HDT) program implemented in three primary schools in a low-income neighborhood in the Netherlands. We find treatment effects on math scores of 0.28 national population standard deviations after one school year ($p < 0.01$). These effects can account for 40% of the math achievement gap between low-income and high-income students in the Netherlands. As most of the evidence on intensive tutoring programs draws on research from the United States, we conclude that (*i.*) HDT programs can be successfully built from the ground up and exported to different institutional settings while maintaining substantial effect sizes, and, (*ii.*) existing income-achievement gaps can be substantially reduced by targeting low-income communities with scalable interventions like HDT.

1. Introduction

Despite decades of comparatively generous welfare state policies in the Netherlands, meaningful achievement gaps between high and low-income families and communities persist and may be widening.[1] Previous policies to address these inequalities have developed largely in a piecemeal fashion and have not been able to structurally reduce these gaps. Recently, however, high-dosage tutoring (HDT) programs have been suggested as a promising way forward to help reducing inequalities in educational outcomes at scale.[2] The popularity of HDT programs might be explained by the combination of two factors: documented sizable treatment effects and (what at least appears from the outside to be) a relatively uncomplicated method of delivery.[3] The latter facilitates scaling-up and exporting the program to other environments, with new and inexperienced people responsible for implementation.

The evidence base for tutoring programs draws mainly on research from low-income settings in the US (e.g. Nickow et al., 2020). It

☆ The authors thank Dinand Webbink and Hessel Oosterbeek for helpful comments and discussions, Isabel Speelman and Shelby Sissing for excellent research assistance, and editorial support. From his role at the University of Amsterdam, one of the authors of this paper (Bowen Paulle) advocated for, and advised on the implementation of the intervention that is investigated in this paper. While retaining his academic position, in September 2021 Paulle became the director of Stichting (Foundation) The Bridge Learning Interventions, a Dutch non-profit implementing high-dosage tutoring. This is a different foundation than the one whose efforts are examined in this paper. The organization that provided the funding for this research was also involved in the funding of the intervention itself. The organization has expressed a desire to remain anonymous.

* Corresponding author.
E-mail address: joppederee@gmail.com (J. de Ree).

[1] See for example Borghans et al. (2018) [in Dutch] and Dutch Inspectorate of Education (2019) [in Dutch]. While achievement gaps may have widened somewhat in the past decade or so, the "big picture" is that achievements gaps between low-income and high-income students are sizable and quite persistent. These patterns are not unique to the Netherlands and can be found all over the world (e.g. Reardon, 2011).

[2] Intensive, or high-dosage, tutoring programs provide tutoring in hourly sessions, for multiple days a week, for an entire school year. It typically involves small group personalized instruction with a student-to-tutor ratio of about 2:1.

[3] See e.g. Guryan et al. (2023) and Kraft (2015), as well as Nickow et al. (2020) and Pellegrini et al. (2021) for overviews of experimental research. In addition to the evidence, policymakers in quite a few countries have picked up on tutoring programs in their efforts to reduce losses due to the COVID-19 pandemic (e.g. the National Tutoring Programme in the United Kingdom).

https://doi.org/10.1016/j.econedurev.2023.102383
Received 22 October 2021; Received in revised form 17 February 2023; Accepted 23 February 2023
Available online 10 April 2023

is not clear how to value these estimates for policymaking decisions in the Netherlands or in other European welfare states. We expect that at least two factors play a role here. First, comparatively high levels of income redistribution and related welfare state provisions exist in the Netherlands, which primarily benefit low-income families.[4] The extent to which tutoring programs are complementary to other state support programs is a priori unclear.[5] Second, scaling up successful programs and exporting it to different contexts tends to reduce effect sizes, because high-quality implementation cannot be maintained consistently.

To investigate the feasibility of substantially reducing the achievement gap between low-income and high-income families, we have cooperated with funders, service providers and schools in an effort to implement an intensive (high-dosage) math tutoring program in all three primary schools in a low-income neighborhood in the Netherlands. The purpose of the intervention was to contribute to a "neighborhood effect", by lifting achievement levels of entire cohorts of students in this low-income area, regardless of their levels of baseline achievement.[6] Administrative data from Statistics Netherlands shows that in these schools, the income level of the median family is at the 15th percentile nationally.[7] The tutoring intervention was modeled after the successful Match Education[8] program which was subsequently developed further by Saga Education.[9] Tutoring is delivered in 2:1 student–tutor ratio and it was intended that students and tutor would work together for the duration of the program. The program provides tutoring in four hourly sessions per week for an entire school year.[10]

To evaluate the effects of the program we used a randomized controlled trial with a roll-out design, where students were randomly assigned to receive tutoring in 4th, 5th or in 6th grade. One-year treatment effects are estimated based on randomly assigned 4th and 5th grade students measured across three cohorts. On average, across cohorts, we document one-year treatment effects of 0.28 national population standard deviations ($p < 0.01$). The one-year effects are qualitatively similar to earlier findings from the US, indicating that exporting the program to a new context did not meaningfully change effect sizes. We do not find evidence for heterogeneous effects across the baseline achievement distribution, indicating that both high and low (baseline) achievers benefited. We also do not find evidence for (positive or negative) effects on reading comprehension.

Our two main contributions are summarized as follows. First, the income-achievement gap in the Netherlands might be significantly reduced by scaling up HDT in primary schools serving low-income communities. Our effect sizes stand out against a comparatively modest, but very persistent income-achievement gap in the Netherlands. Using administrative data we show that an effect size of 0.28 roughly corresponds to the difference in math achievement between low-income Dutch primary students (at the 10th percentile of the parental income distribution) and median income Dutch primary students. Over the past few years, new, and similar high-dosage math tutoring programs have been introduced in other low-income areas in the Netherlands.[11] These efforts to scale HDT, however, remain for now at a limited geographical scale (i.e. other neighborhoods). Scaling even further would involve considerable challenges with regard to funding and implementation.

Second, we demonstrate that HDT can be exported to very different institutional settings while maintaining meaningful effect sizes. The project examined here was implemented at a pilot scale, but the fact that it was successfully built from the ground up underscores the scalability of these programs.[12] A benefit of the tutoring program is that it runs largely independently of regular school operations and that it does not require much behavioral change from classroom teachers. Kraft (2020) argues that "the challenge posed by taking programs to scale is largely proportional to the degree of behavioral change required to implement a program". Approaches that focus on improving teacher effectiveness, for example, might not be as easily exported or scaled up (see e.g. Jacob & Lefgren, 2004).

The central importance of scalability also comes into view when considering the enormous tutoring projects that are, or might soon be used in attempts to reverse achievement losses caused by the COVID-19 pandemic.[13] These large scale interventions exemplify a belief in the scalability of tutoring programs and our research provides a new reference for this. Recent research has estimated the effects of school closures in the Netherlands at 8% of a standard deviation, up to 11% for disadvantaged youth (Engzell et al., 2020). Such losses can be addressed by the HDT model that we study in this paper. A quick recovery of these losses might be important as some research finds long term effects of reduced time in school (Andrabi et al., 2021).

The remainder of the paper is organized as follows. In Section 2 we introduce and describe the HDT intervention. In Section 3 we present the research design and in Section 4 we present and discuss our findings. In Section 5 we use administrative data from Statistics Netherlands to estimate the relationship between primary student achievement and parental income, as a benchmark for interpreting our findings. Section 6 concludes and suggests areas for further research.

2. The high-dosage tutoring program

The high-dosage math tutoring program was introduced in 2015/16 in all three primary schools in a low-income neighborhood in the Netherlands.[14] The program was part of a broader effort to support children of low-income families. It was built from the ground up in a collaborative effort by schools, funders and service providers who had no prior experience with implementing tutoring interventions. In large part, the tutoring program aimed at replicating the successful Saga Education tutoring program. Staff from Saga Education have also served as consultants on this project. Other than temporary consulting grants, however, there was no formal relationship between the Dutch consortium and Saga Education. When the Dutch HDT program was being developed, existing programs were starting to show promising impacts in low-income settings in the US, particularly for young adolescents (e.g. Cook et al., 2014 and Fryer Jr., 2014 as well as preliminary

[4] See e.g. Alesina and Glaeser (2005) who document differences in service provision between European countries and the US.

[5] Jackson et al. (2016) for example documents evidence on the importance of diminishing returns to school spending. (Dynamic) complementarities however have also been shown. For example, Johnson and Jackson (2019) show that the benefits of Head Start were larger when followed by access to better funded schools.

[6] While there was some variation with regard to the socioeconomic status of the families of children in these schools, the guiding assumption was that offering access to HDT to all children in these schools, would automatically mean reaching predominately children of low-income families.

[7] The 10% highest income earning families in this neighborhood have income levels around the national median. In other words, the income distribution in this neighborhood, roughly spans the bottom half of the national parental income distribution.

[8] See https://www.matcheducation.org/export/prior-projects/district-partnerships/ and Kraft (2015) for background information on the Match Education HDT program as well as Fryer Jr. (2014) who analyzes best practices of charter schools in the US, including high-dosage tutoring.

[9] https://www.sagaeducation.org/our-story

[10] The Dutch program provides tutoring for four days a week, while the original Match model provides tutoring on all five school days.

[11] One example of this is a high-dosage math tutoring project in primary education in Amsterdam (De Ree & Paulle, 2021).

[12] See also Davis et al. (2017) for a discussion on the economics of scale-up, and Kraft (2020) for arguments about the importance of scalability when assessing educational interventions.

[13] See e.g. the National Tutoring Programme in the UK or tutoring initiatives in the US as part of the so-called American Rescue Plan.

[14] As some stakeholders have expressed a desire to remain anonymous, we do not mention some of the local details of the HDT intervention.

findings from ongoing research projects). By now, the Saga Education program has been successfully RCT-tested multiple times and efforts to scale up this program are currently underway (Guryan et al., 2023). The scaling-up process itself is also being investigated (Davis et al., 2017).

The development of HDT programs was informed by the insight that intensive individualized instruction can be very effective, but also very costly (Bloom, 1984). HDT programs therefore take elements of individualized instruction, but limit the cost of implementation by providing tutoring for (only) one-hour sessions each day and by relying on paraprofessional tutors (rather than fully certified teachers). The Dutch program provided tutoring for an entire school year with four hourly sessions per week. The tutoring was delivered by teams of six tutors and a so-called site director. Recruitment of tutors was based on responses to job postings, leading to job interviews. A minimum education was not required for aspiring tutors. In order to be hired, they had to pass a standardized math test at the 2F level[15] and an assessment. Tutor pay was comparable with that of classroom assistants. For the period we study in this paper the tutors were typically recent graduates of BA or MA programs with limited work experience. Very few (if any) had studied mathematics at the BA or MA levels and/or were certified to teach. Some had worked (informally) as tutors, but none of them had any previous experiences with this specific tutoring approach. As with the Saga program, these tutors received substantial on-the-job training and coaching by the site director. Site directors were expected to be on site at all times. The on-the-job training and coaching of tutors is an important element of this program as tutors were typically taking this job for only one or two years. The site directors of this program received some training from Saga Education.

In the beginning of the school year the service provider, informed by the teachers, arranged participating students in pairs and matched each pair to a tutor. The formation of these pairs was based on perceived fit and on math achievement levels at baseline. The intention was that pairs of students and their tutor would work together for an entire school year. Tutoring was implemented during regular school hours and participants therefore missed four hours of normal classroom activity. It was agreed that two of the four weekly sessions replaced regular classroom math instruction and practice, while the remaining two sessions would not replace any core subjects, such as math, reading or spelling. It was also agreed that classroom teachers would only introduce new concepts when all students (i.e. also those participating with HDT) were present. Morning and afternoon sessions were alternated so that participants would have their tutoring sessions (more or less) balanced across mornings and afternoons.

The intervention was meant to facilitate practice with mathematical concepts that were previously introduced by the classroom teacher (recently or years ago). Tutors would spend about 20% of the time reviewing, explaining or demonstrating concepts. The other 80% of the time, students would themselves be working on math problems under direct supervision of their tutors. During the two hours that tutoring replaced regular math instruction, the tutors would explicitly follow the textbook that was used in class. The site director would coordinate with the classroom teacher about the topics that needed to be covered. Tutors (via the site director) subsequently received instructions about which problem sets were to be discussed. This way, participants would essentially cover the same material as the nonparticipating (control) students. During these two hours the flexibility to personalize instruction to individual needs was limited. However, working on the curriculum also helped tutors to explore knowledge gaps.

During the other two weekly sessions, tutors would personalize instruction to the needs of the individual student. These two "free" hours, therefore, were truly additional hours dedicated to math practice.[16] Based on a continuing assessment of a student's need, tutors would provide additional explanation and practice. The tutoring program did not (at least at the time) have a clear curriculum which would provide specific direction to the sessions. The assessment of knowledge gaps and figuring out ways of helping students, therefore, might have depended to some extent on the creativity of individual tutors. However, challenges that tutors were facing were meant to be discussed under supervision of the site director during (collective) lesson planning sessions. Lesson planning would take place in the afternoon, after tutoring.

Aside from the explicit focus on math practice, the tutoring program also had a (somewhat implicit) socioemotional component. Tutoring provides a stable environment in which two students and a tutor work together on an almost daily basis. In this setting there are opportunities to quickly develop a connection and get to know each other. This was meant to create an environment in which students felt comfortable making mistakes and trying out new things. Experiencing progress might generate a sense of self-confidence, which might be important especially for students who are behind. Socioemotional skills, such as those associated with confidence or motivation, are likely important factors in explaining success in school and in life more generally (Heckman et al., 2006). The tutoring intervention had aspects of mentoring relationships that have been found to boost socioemotional skills (e.g. Kosse et al., 2020). To expand the support network around these children tutors also involved parents by contacting them once per week.

Table 1
5 clusters from which students were assigned to 5 HDT sessions.

	4th Grade	5th Grade
Cluster 1 [School A]	Class 4	Class 5
Cluster 2 [School B]	Class 4a	Class 5a
Cluster 3 [School B]	Class 4b	Class 5b
Cluster 4 [School C]	Class 4a	Class 5a
Cluster 5 [School C]	Class 4b	Class 5b

3. Design and data

The intervention was evaluated using a randomized controlled trial with a rollout design, where the timing of receiving tutoring was determined by chance. It was agreed with the schools that all students of the eligible cohorts would receive tutoring during one school year, either in 4th, 5th or in 6th grade. The starting point of the random assignment of students was an operational restriction. A team of six tutors were hired to deliver the program to the selected students in five daily sessions, of one class hour per session. In each session, therefore, there were 12 available seats (six tutors, with two students each).

The five sessions were then distributed across the three participating schools (roughly) in proportion to the size of the school. School A has one class per grade and was assigned one (of five) daily sessions. The other two schools (B and C) each have two classes per grade and were assigned two daily sessions each. Within schools, clusters of one 4th grade and one 5th grade class were formed where one daily session was assigned to one cluster. Table 1 presents the structure of these clusters and the way they are distributed across the participating schools.

The tutoring sessions were filled with randomly selected 4th and 5th graders from the corresponding cluster. Random assignment of students was done in the beginning of three school years (2015/16, 2016/17 and 2017/18). The primary objective of the evaluation was

[15] https://www.rijksoverheid.nl/onderwerpen/taal-en-rekenen/referentiekader-taal-en-rekenen for reference levels in math and literacy. The Dutch *2F* level is comparable to international ISCED level 2.

[16] The fact that tutoring took place in part during regular math instruction and practice and was only partly personalized, is one way in which the Dutch model is an adaption of the Saga model. In the Saga model, all sessions are additional to the regular math curriculum.

Table 2
Number of students assigned to the $T5-C5$ and $T4-C4$ conditions, by cluster, by school year.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	School year 2015/16				School year 2016/17				School year 2017/18			
	$T4$	$C4$	$T5$	$C5$	$T4$	$C4$	$T5$	$C5$	$T4$	$C4$	$T5$	$C5$
Cluster 1: School A	1	22	11	11			12	11			12	12
Cluster 2: School B a	4	17	8	8	4	18	8	9			10	12
Cluster 3: School B b	2	18	10	10	3	17	9	9	9	9	4	5
Cluster 4: School C a			12	13	3	17	9	9	4	4	8	7
Cluster 5: School C b	3	12	9	9	6	5	6	7	9	9	2	3
Total	10	69	50	51	16	57	44	45	22	22	36	39

Table 3
Treatment across conditions.

	4th Grade	5th Grade	6th Grade
A: 5th grade experimental sample			
$T5$	Not treated	**Treated**	Already treated
$C5$	Not treated	Not treated	**Treated**
B: 4th grade experimental sample			
$T4$	**Treated**	Already treated	Already treated
$C4 \rightarrow T5$	Not treated	**Treated**	Already treated
$C4 \rightarrow C5$	Not treated	Not treated	**Treated**

to estimate the one-year treatment effect for 5th grade. This objective was formalized by the prioritization of 5th graders in the assignment to treatment. 4th grade students were only added in case the assignment of 5th graders did not reach the required 12 students per session. The random assignment of students was operationalized as follows:

1. In the beginning of each school year, 4th and 5th grade students in the three participating schools were listed, by cluster-class (see Table 1).
2. Based on the list, eligibility was determined. 4th and 5th grade students who were randomly assigned to treatment in an earlier school year, were excluded from the sampling frame. (This was mainly an issue for 5th grade students in year t who had been randomly assigned to treatment in 4th grade in year $t-1$.)
3. In a first step, eligible 5th grade students were randomly assigned to a treatment ($T5$) and a control ($C5$) condition using a stratified randomization procedure.[17] An effort was made to maintain roughly equal sample sizes across the $T5$ and $C5$ conditions within each stratum. Students assigned to $T5$ would receive treatment in 5th grade. The remainder, the students assigned to $C5$, would receive treatment in 6th grade. The combined set of students assigned to $T5$ and $C5$ are referred to in the paper as the **5th grade experimental sample**. Denote $N_{k,T5}$ as the number of students assigned to $T5$ in cluster k.
4. If $N_{k,T5} < 12$, additional 4th grade students from cluster k would be also randomly assigned to a treatment ($T4$) and a control ($C4$) condition such that in each cluster the total number of students assigned to $T4$ and $T5$ would add up to 12 ($N_{k,T5} + N_{k,T4} = 12$). Students assigned to $T4$ would receive treatment in 4th grade. The combined set of students randomly assigned to $T4$ and $C4$ are referred to in the paper as the **4th grade experimental sample**.

Table 2 presents the number of students assigned to each condition, by cluster and by school year of random assignment. The number of students assigned to $T4$ is typically much smaller than the number of students assigned to $T5$. Also, for four different cluster–year pairs 4th graders were not randomly assigned at all, because the cluster–year already provided enough 5th graders to reach the required 12 students.

Table 3 shows the exact treatment-control contrast, for the 4th and the 5th grade experimental samples. For the 5th grade experimental sample the HDT program starts in 5th grade, for the $T5$ selection. Students assigned to $C5$ are treated in 6th grade.[18] For the 4th grade experimental sample, the HDT program starts in 4th grade for the $T4$ selection. In 5th grade, students that were previously assigned to $C4$ are eligible for random assignment in 5th grade. Those who are assigned to $C4$, and then to $T5$, receive treatment in 5th grade. Those assigned to $C4$, and then to $C5$, receive treatment in 6th grade. Note that the number students assigned to $C4$ and then to $T5$ is only a subset the students who are assigned to $T5$.

Table 3 shows that all students receive math tutoring at one point in their primary school careers. The design therefore does not allow for the estimation of longer term effects. Instead, the design allows for estimating one-year treatment effects, by comparing $T4$ to $C4$ at the end of 4th grade, and by comparing $T5$ to $C5$ at the end of 5th grade. It also allows for estimating the differences between receiving tutoring early or later, for example, by comparing outcomes between $T5$ and $C5$ in 6th grade, or by comparing outcomes between $T4$ and $C4$ in 5th and/or in 6th grade. In principle it is also possible to estimate fade-out effects, or decay, by comparing $T4$ to the $C4 \rightarrow C5$ selection in 4th and in 5th grade. In the results Section 4.1 we present main estimates of the one-year treatment effects. In Section 4.2 we present results from follow-up measurements. Note that there is considerable overlap between students assigned to $C4$ in school year t and the 5th grade experimental sample $(T5, C5)$ in the next school year $t+1$. However, because this overlap is not perfect we have decided to present separate analyses for the 4th grade and the 5th grade experimental samples.[19]

3.1. Outcome data

Our main outcome variables are scores on comparable standardized math and reading comprehension tests developed by Dutch test developer Cito. The tests are used by a large share of primary schools in the Netherlands.[20] Schools rely on these tests for monitoring achievement

[17] Students were stratified by cluster, school year and class. For the 2017/18 school year, students were also stratified based on baseline math test score.

[18] For the tutoring of 6th graders, additional tutors were hired.

[19] The lack of overlap between $C4$ and the 5th grade experimental sample in the year after, concerns for example the entire 2015/16 cohort of the 5th grade experimental sample, as well as cluster–years for which 4th graders were not randomly assigned.

[20] https://cito.com/student-tracking-systems

of each student over time and for comparing students to the national distribution of achievement of each age group. Students are typically tested twice a year. Semester-to-semester performance on these tests are used as a basis for secondary school track assignment at the end of primary education. The tests, therefore, are high stakes.

In the analysis we standardize test scores with respect to the (estimated) national means and (estimated) national standard deviations for each grade-semester. That is, we scale the test scores Y_{igs}, for student i in grade g in semester s as follows:

$$y_{igs} = \frac{Y_{igs} - \mu_{gs}}{SD_{gs}} \tag{1}$$

The test developer Cito provides survey estimates of national means μ_{gs} and quintile cutoffs for each grade-semester. Based on the quintile cutoffs we estimate SD_{gs} by dividing the distance between the 20th and the 80th percentile by 1.683, which is the distance measured in standard deviation units between the 20th and the 80th percentile of a normal distribution. Cito reports show that the distribution of scores on these tests in the general population is approximately normal (Cito, 2015) such that our approach to scaling seems justified. In the appendices we also present results based on test scores that are scaled with the standard deviation of the control group. As an alternative math outcome we also use scores on the so-called math speed test (*tempotoets rekenen*). For the speed test, students would return as many correct answers as they can in a short, pre-specified period of time (e.g. 5 min).

In addition we use data from a teacher questionnaire. It includes a total of 58 questions and statements about the student, measuring concepts as student behavior, teacher–student relationships, teacher perceived prosocial behavior and self-confidence. At the end of the academic year, teachers would fill out the questionnaire for each student in their class. The questionnaire is based on a questionnaire from the Dutch education cohort studies PRIMA and COOL (e.g. Jungbluth et al., 2001 and Driessen et al., 2009). The reliabilities (Cronbach's alpha) of the scales range from 0.48 to 0.85 in our data (see Appendix I for the complete questionnaire [translated from Dutch to English]).

In Appendix A we show baseline summary statistics, including balance tests. The tables show that students assigned to treatment and control conditions are similar on observable baseline characteristics. In Appendix E we present tests on random attrition.

3.2. Statistical models

The one-year treatment effects are estimated using the following linear regression model.

$$y_{i,t,s} = \alpha_{t,s} + \beta_{t,s} T_i + \gamma_{0,1} y_{i,0,1} + \gamma_{0,2} y_{i,0,2} + \sum_{b=1}^{B} I\left(\text{stratum}_b = 1\right) + u_{i,t,s} \quad \forall t \geq 1 \tag{2}$$

for student i observed in semester s of school year t after random assignment. The model is estimated on the 4th grade and 5th grade experimental samples as well as pooled. For the 4th grade experimental sample $T_i = 1$ if $T4_i = 1$, while students assigned to $C4$ are the omitted category. For the 5th grade sample $T_i = 1$ if $T5_i = 1$, while students assigned to $C5$ are the omitted category. The variables $y_{i,0,1}$ and $y_{i,0,2}$ are baseline outcome scores, measured at the end of the first and second semester of the school year prior to random assignment.[21] The variables $I\left(\text{stratum}_b = 1\right)$ are stratum fixed effects.

If we pool data across years and/or experimental samples, we allow all parameters, except for the causal parameter $\beta_{t,s}$, to be different between the years of random assignment (2015/16, 2016/17 and 2017/18) and between the 4th and 5th grade experimental samples. In the regressions we use weights based on the propensity score. Weights for treatment observations are $1/p$ and weights for control observations are $1/(1-p)$ where p is the probability of assignment to treatment. The weighting accounts for differences in the selection probabilities across strata, so that our estimates reflect the average treatment effect across the entire experimental sample(s). If we pool data, we compute clustered standard errors at the level of the student. Students might appear twice in the data, as part of the 4th grade sample and as part of the 5th grade sample.

Table 3 shows that for the 4th grade experimental sample we could compare outcomes across three different groups. In Section 4.2 we use the following augmented regression model on our 4th grade experimental sample:

$$\begin{aligned} y_{i,t,s} = {} & \alpha_{t,s} + \beta_{t,s}^{T4} 1\left(T4_i = 1\right) + \beta_{t,s}^{C4 \to T5} 1\left(C4_i = 1, T5_i = 1\right) \\ & + \gamma_{0,1} y_{i,0,1} + \gamma_{0,2} y_{i,0,2} + \sum_{b=1}^{B} I\left(\text{stratum}_b = 1\right) + u_{i,t,s} \quad \forall t \geq 1 \end{aligned} \tag{3}$$

where $1\left(T4_i = 1\right)$ is a dummy variable for students assigned to $T4$ and $1\left(C4_i = 1, T5_i = 1\right)$ is a dummy variable for students assigned to $C4$ and then to $T5$. The omitted category are students assigned to $C4$ and then to $C5$. The parameter $\beta_{t,s}^{T4}$ compares students assigned to $T4$ with students assigned to $C4 \to C5$. Because students assigned to $C4 \to C5$ only receive treatment in 6th grade, we can use 4th grade outcomes to estimate one-year treatment effects and 5th grade outcomes to measure fade-out/decay.[22] While in principle the estimation of treatments effects and fade-out is possible, the 4th grade effects are imprecisely estimated.

4. Results

4.1. Main results: one-year treatment effects

Table 4 presents the half-year and one-year treatment effects for math, pooled across the three years of implementation.[23] For Table 4A we have pooled the data across the 4th and 5th grade experimental samples. The treatment effects are precisely estimated with t statistics of around 4 (for the column [2] results). We estimate half-year treatment effects of 0.18 national population standard deviations and one-year treatment effects of 0.28 national population standard deviations. In Table 4B and 4C we present separate results for 5th grade and 4th grade experimental samples. The 5th grade results are more precisely estimated than the 4th grade results. The lack of precision for 4th grade are due to the smaller sample sizes in the treatment condition $T4$. The magnitude of the 4th and 5th grade results however

[21] For our main outcomes we typically use two baseline scores, while for the math speed test and the survey we only observe a single baseline outcome score. If one of the two baseline scores is not observed, we impute the missing score with a prediction out of the non-missing baseline score, or otherwise out of test scores that were observed two years prior to randomization. For prediction we use a simple linear OLS regression model on the nonmissing data. If, for any student, no prior test score data is observed, we set both baseline scores to zero and include an additional dummy variable in the regression model that is 1 for observations with unobserved baseline outcome data, and 0 otherwise. See e.g. De Ree et al. (2018) who deal with missing baseline data in the same way. For precise estimation of the causal parameters, it is important to control for baseline values in our setting. In Appendix B we show that our main results are robust to variations in the exact way of controlling for baseline values.

[22] When estimating Eq. (3) on the 4th grade experimental sample, we essentially compare $T4$, $C4 \to T5$ and $C4 \to C5$. For this model, we derive the probabilities of selection into each of these three groups. The weights we use in the regressions are one over the respective group-specific selection probabilities.

[23] Estimated effects for different cohorts are not significantly different from each other. $p = 0.44$ and $p = 0.67$ for tests on equal half-year and one-year treatment effects across the three cohorts of the 5th grade sample. $p = 0.42$ and $p = 0.54$ for tests on equal half-year and one-year treatment effects across the three cohorts of the 4th grade sample.

Table 4
Treatment effects on math scores, measured in national population standard deviation units.

	(1) Half-year	(2) one-year
A: Pooled 4th and 5th grade samples (3 cohorts)		
Treatment effect	0.18***	0.28***
	(0.06)	(0.07)
	[441]	[434]
B: 5th grade experimental sample (3 cohorts)		
Treatment effect	0.25***	0.28***
	(0.06)	(0.06)
	[255]	[251]
C: 4th grade experimental sample (3 cohorts)		
Treatment effect	0.09	0.27*
	(0.11)	(0.15)
	[186]	[183]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Estimated treatment effects are measured in national population standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

Table 5
Treatment effects on reading comprehension scores, measured in national population standard deviation units.

	(1) Half-year	(2) One year
A: Pooled 4th and 5th grade samples (3 cohorts)		
Treatment effect	−0.07	−0.07
	(0.07)	(0.08)
	[444]	[409]
B: 5th grade experimental sample (3 cohorts)		
Treatment effect	−0.09	−0.08
	(0.10)	(0.10)
	[256]	[222]
C: 4th grade experimental sample (3 cohorts)		
Treatment effect	−0.05	−0.06
	(0.11)	(0.12)
	[188]	[187]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Estimated treatment effects are measured in national population standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

are similar. We also cannot reject equal effects across the 4th and 5th grade experimental samples ($p = 0.20$ for a test on equal half-year effects and $p = 0.95$ for a test on equal one-year effects).

In Appendix C we present the estimated treatment effects measured in control group standard deviation units (i.e. Glass's Δ). The one-year treatment effect using pooled 4th and 5th grade data is 0.26 control group standard deviations. This indicates that the spread of achievement within our (control group) sample is marginally greater than the spread of achievement in the population of all primary students in the Netherlands. While both results are relevant, the results expressed in national population standard deviations allow for a more direct comparison to policy objectives. In Appendix C we also report the treatment effects on the alternative math outcome: the math speed test. Pooled across the 4th and 5th grade experimental sample, we estimate treatment effects on the speed test data of 0.42 control group standard deviations.

At the mean, a treatment effect of 0.28 standard deviations corresponds to approximately one decile. Kraft (2020) qualifies such effects as large when compared to other educational interventions. Nickow et al. (2020) investigate 96 randomized studies of tutoring interventions in math and in literacy. They find an average effect size of 0.37 across these studies. However, precisely comparing effect sizes across different settings and tests is challenging. (For example, in our study we find effects of 0.42 control group standard deviations on the math speed test and 0.26 control group standard deviations on the Cito math test.) For policymaking in the Netherlands, and in comparable settings perhaps, the treatment effects presented in Table 4 seem most relevant. These effects are based on a general standardized achievement test and they allow for a direct comparison against the national distribution of achievement of the relevant age group. In Section 5 we compare these treatment effects against the math achievement gap between low-income and high-income students in the Netherlands.

As the intervention focused on math and was conducted during school hours, it might have negatively affected achievement in other domains. In Table 5 we therefore present treatment effects based on reading comprehension tests. While the point estimates are negative, the results do not suggests strong evidence for negative spillovers. Future research could pool data from multiple experiments to increase power to detect such unintentional effects. Guryan et al. (2023) also do not find effects on reading.

4.2. Exploring results from follow-up measurements

The rollout design allows for evaluating whether the timing of receiving treatment matters. Two aspects might simultaneously play a role here. First, treatment effects of the same intervention might differ between age groups. Research suggests that the effects of most educational interventions tend to decline as the age (or grade level) of the target population increases (see e.g. Cascio & Staiger, 2012 for a summary). Second, test score effects of educational interventions tend to fade out over time (see e.g. Cascio & Staiger, 2012 for a summary).

In Table 6A we compare outcomes between $T5$ (treatment in 5th grade) and $C5$ (treatment in 6th grade) across time. For this we use regression Eq. (2) on the 5th grade experimental sample. We focus on the 5th grade cohorts 1 and 2, the cohorts for which we have access to follow-up data from 6th grade. Column (1–2) present the half-year and one-year treatment effects pooled across the cohorts 1 and 2.[24] In 6th grade $C5$ (the omitted category in the regression model) received treatment. The parameter on $T5$ in column (3) is negative, indicating that students assigned to $T5$ score lower on the test than students assigned to $C5$, halfway into 6th grade. This estimate is not statistically significant. Unfortunately we do not have test score data to compare $T5$ and $C5$ at the end 6th grade, when both $T5$ and $C5$ had one full year of tutoring.[25]

In Table 6B we compare outcomes between $T4$ (treatment in 4th grade), $C4 \rightarrow T5$ (the subset of $C4$ that receives treatment in 5th grade) and $C4 \rightarrow C5$ (the subset of $C4$ that receives treatment in 6th grade) across time. For this we use regression Eq. (3) on the 4th grade experimental sample. We focus on the cohorts 1 and 2 of the 4th grade experimental sample for which we have complete data until 5th grade. The parameter on $T4$ estimates the half-year and one-year treatment effects in the columns (1–2), by comparing $T4$ to $C4 \rightarrow C5$ (the omitted category). The parameters are somewhat imprecisely estimated just like the estimates presented in Table 4C. The fact that the parameter is decreasing in magnitude right after the first year, suggests fade-out effects. However, we cannot formally reject the absence of fade-out from this pattern alone ($p = 0.33$).[26]

[24] Table 4B presents estimates pooled across all three cohorts.

[25] The math tests we use are not administered in the second semester of 6th grade.

[26] To test no fade-out, we test equality of the parameter on $T4$ of column (2) and (4) of Table 6B.

Table 6
Follow-up measurements for math scores, measured in national population standard deviation units.

	(1) 5th grade	(2)	(3) 6th grade	(4)
	sem. 1	sem. 2	sem. 1	sem. 2
A: 5th grade sample (cohort 1&2)				
β	0.25***	0.26***	−0.08	
	(0.07)	(0.07)	(0.09)	
	[183]	[181]	[169]	
	4th grade		5th grade	
	sem. 1	sem. 2	sem. 1	sem. 2
B: 4th grade sample (cohort 1&2)				
β^{T4}	0.09	0.23	0.16	0.08
	(0.14)	(0.19)	(0.16)	(0.11)
$\beta^{C4\rightarrow T5}$	0.00	0.07	0.43***	0.32***
	(0.10)	(0.10)	(0.14)	(0.11)
	[136]	[135]	[133]	[128]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Regression estimates are based on equation (2) for panel A and based on equation (3) for panel B. Estimated parameters are measured in national population standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

A comparison between the parameter on $T4$ and the parameter on $C4 \rightarrow T5$ at the end of 5th grade can be used to evaluate whether the timing of receiving tutoring matters. At the end of 5th grade, students who have received tutoring in 5th grade (as opposed to 4th grade) score significantly higher ($p = 0.04$).[27] This indicates that receiving treatment later (in 5th grade instead of in 4th grade) yields better outcomes at the end of the cycle, when all students have received treatment.

Taken together, the results provide some evidence that students who receive treatment later are better off (at least when measured right at the end of the cycle) and that, potentially, some of this is due to fade-out effects. The sample sizes however, particularly of those assigned to $T4$, are insufficient to generate clear statistical support for the fade-out hypothesis. Existing research however suggests that fade-out is common and our data seems consistent with this result (Cascio & Staiger, 2012). The presence of fade-out effects however does not rule out longer term effects. This might be true in particular for interventions with a strong socio-emotional component, see e.g. Deming (2009), Chetty et al. (2014) and Sorrenti et al. (2020). Measurable short term gains might go unnoticed in the medium term as they manifest in (skill) domains that are not easily measured. Key aspects of the tutoring intervention are small groups, stability, and individualized instruction. Within this setting, students repeatedly experience and share successes. This might influence levels of self-confidence. Based on our design however we cannot draw any conclusions about these longer term effects.

4.3. Exploring heterogeneous effects

In this section we explore the heterogeneity of the treatment effects across the baseline distribution of achievement and for different levels of parental education. All students of the cohorts involved would participate with the HDT program. Because most of the variation in achievement is within communities rather than between them, we have many low-income, but high-achieving students in our data. In Fig. 1 we investigate the heterogeneity of the treatment effects by baseline achievement for the 5th grade experimental sample, the subset of the data for which we obtain precisely estimated treatment effects. Because HDT offers personalized instruction one might not expect that treatment effects differ much across the baseline distribution of achievement. The literature however shows mixed results with regard to the potential heterogeneity of treatment effects (see e.g. Kraft & Falken, 2021 for references).

We construct a change score between the baseline (before treatment) and the endline (after one year of treatment) for the 5th grade experimental sample, and plot these against percentiles of the baseline score. We fit curves for the $T5$ and $C5$ selections using a local linear polynomial smoother.[28] The curves are slightly downward sloping, consistent with some regression to the mean. The figure however does not indicate any clear heterogeneity in the treatment effects. Using randomization inference we also cannot reject the null of a constant effect ($p = 0.41$).[29]

The results indicate that high-achievers and low-achievers benefit from the additional support in a similar way. In disadvantaged communities, high-achieving students might face very particular challenges as they might not always be able to rely on their parents for help with math problems for example. Note that about one-third of our sample has parents with low to very low levels of education.[30] It is plausible that investments in these high-potential, disadvantaged students may yield important (economic) returns in the long run. Our findings also suggest that targeting based on baseline achievement is not specifically warranted from the point of view efficient public spending.

Another feature of the intervention was the targeting of low-income communities without specifically distinguishing relatively high-income and low-income families within these communities. This somewhat global targeting of communities might raise efficiency concerns as governments (or others) have only limited means to support families this way. Decisions on allocations of resources could also depend on relative returns. In Appendix H we test for heterogeneous treatment effects for different levels of parental education within our sample. We do not find statistically significant differences between students with different levels of parental education. This result suggests that targeting low-income communities as a whole does not bear clear risks of inefficient public spending. Instead, by targeting low-income communities as a whole one might benefit from economies of scale in the implementation.

4.4. Survey results

In this section we investigate the effects of the tutoring program on outcomes of a teacher's survey. In Appendix I we list the 58 items of this questionnaire. Classroom teachers fill out this questionnaire for each student at end of the school year. The survey would measure some socioemotional skills as well as concepts like perceived achievement, student behavior and teacher–student relationships. Out of the 58 items 10 composite scores were constructed. The composite scores were standardized using the standard deviation in the control group. Generally, the students in our sample were also scored in the year prior to the intervention. We incorporate this baseline score as a control variable in the regressions. While we had the opportunity to use this survey data, the survey was not administered with the purpose of evaluating the effects of the tutoring program.

Table 7A presents the pooled estimates on the 10 composite scores. The results show that for most outcomes the treatment effects are not statistically significant. We see positive effects however on (teacher

[27] We test equality of the parameter on $T4$ of column (4) and the parameter on $C4 \rightarrow T5$ of column (4) of Table 6B.

[28] To fit the polynomial we use the LPOLY command in Stata, using a bandwidth of 10 percentile points.

[29] The procedure for estimating the test statistic is explained in the notes below Fig. 1. Appendix D presents these figures separately by cohort. We find consistently that treatment effects do not seem to differ between high and low baseline achievement.

[30] Both parents of about 1/3 of our sample have lower pre-vocational secondary education or less (see Appendix A).

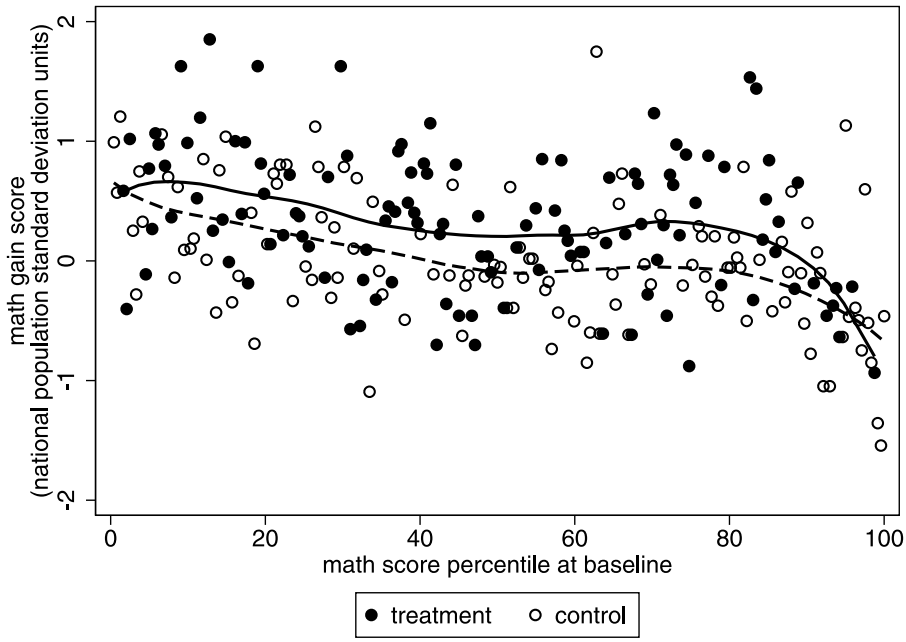

Fig. 1. Comparing gain scores in math between treatment $T5$ [solid curve] and control $C5$ [dashed curve], as a function of baseline math score percentiles.
Notes: We test whether the treatment effect depends on the baseline outcome percentile using a randomization inference procedure. The procedure followed the following steps. *1.* We start by computing $\sqrt{\sum_p \left(\tau(p) - \overline{\tau}(p)\right)^2}$, the test statistic of interest. This is the square root of the mean squared difference between the percentile p specific estimated treatment effects and the (unweighted) mean. *2.* In the same way as the original randomization was done, we reassigned treatment and control 5,000 times. For each draw, we compute the same root mean squared error. *3.* We compare the test statistic of interest against the distribution of 5,000 pseudo test statistics. The assessment indicates a p-value of 0.48. Hence, we cannot reject the null that the treatment effect is independent of baseline achievement.

Table 7
Treatment effects on survey results, measured in control group standard deviation units.

	(1) Achievement	(2) Student behavior	(3) Teacher student relationship	(4) Parental involvement	(5) Academic development	(6) Prosocial behavior	(7) Self confidence	(8) Ambitious	(9) Stress	(10) Organized
A: Pooled 4th and 5th grade experimental samples										
Treatment effect	0.26**	0.19**	0.09	−0.02	0.24**	0.01	0.05	0.01	0.03	−0.04
	(0.11)	(0.09)	(0.10)	(0.08)	(0.10)	(0.09)	(0.09)	(0.10)	(0.12)	(0.09)
	[392]	[392]	[392]	[392]	[392]	[392]	[392]	[392]	[392]	[392]
B: 5th grade experimental sample										
Treatment effect	0.42***	0.13	0.09	0.01	0.33***	0.03	−0.16	0.05	0.05	−0.08
	(0.14)	(0.10)	(0.11)	(0.11)	(0.12)	(0.12)	(0.13)	(0.12)	(0.14)	(0.13)
	[208]	[208]	[208]	[208]	[208]	[208]	[208]	[208]	[208]	[208]
C: 4th grade experimental sample										
Treatment effect	0.09	0.24	0.09	−0.05	0.14	−0.02	0.28**	−0.02	0.01	−0.00
	(0.18)	(0.15)	(0.16)	(0.10)	(0.17)	(0.13)	(0.13)	(0.18)	(0.20)	(0.14)
	[184]	[184]	[184]	[184]	[184]	[184]	[184]	[184]	[184]	[184]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Estimated treatment effects are measured in control group standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

perceived) student achievement, (teacher perceived) student behavior and (teacher perceived) academic development. In Table 7B and 7C we present separate results for the 5th and 4th grade experimental samples. This shows that for 5th grade, the results in column (1) and (5) align well with the strong results on test scores presented in Tables 4 and 13. These results turn out to be mainly driven by effects on similar survey items. We find are negative effects on statements like "this child is underperforming" [contributing to column (1)] and "this child can do better than he/she is right now" [column (5)] and by positive effects on statements like "I am satisfied with this child's academic performance" [column (5)]. The positive estimates presented in columns (1) and (5) therefore seem to indicate that teachers notice the improved performance of treated students. This confirms the relevance of our main findings.

The significant results presented in column (2) also suggest that teachers notice improved classroom behavior after tutoring. The positive effects are mainly driven by positive effects on items stating that students follow class rules, work precisely, and get along well with classmates. For other outcome measures, such as teacher–student relationships, parental involvement, and ambition and stress, the estimated effects are small and not statistically significant. For self-confidence, the significant parameter we find on the 4th grade sample is not robust.

The results taken together suggest that the effects of the intervention are observable by teachers. Students do better in class and teachers

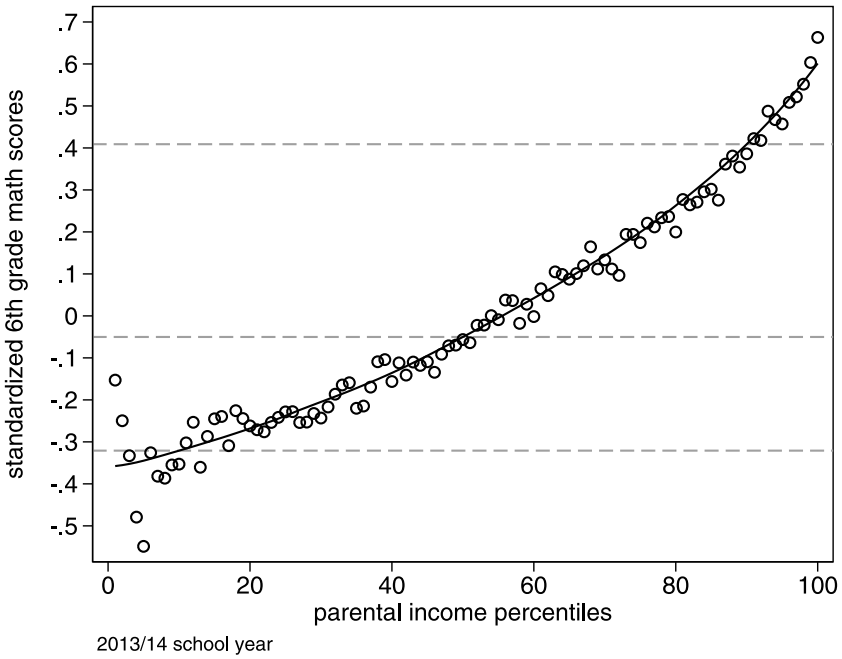

Fig. 2. Fitted relationship between average math scores at the end of primary education and parental income (in percentiles).
Notes: The curve represents the relationship between parental income and the average math scores of primary students at the end of 6th grade. The dots are mean scores for each parental income percentile. The vertical dashed lines represent the conditional mean math scores for the 10th, 50th and 90th percentile of the parental income distribution. The distance between the scores associated with the 10th percentile and the median is 0.27 national population standard deviations. This estimate depends only weakly on the year of measurement and is roughly equal to the estimated one-year treatment effects of the HDT program.

see this. Perhaps as a consequence, students might also show more productive classroom behavior. The effects are broadly in line with the positive treatment effects we find on math achievement. We are not confident in making strong statements about our null effects, even though some of them are reasonably precisely estimated. These findings should also be viewed in the light of the limitations of survey-based instruments. It is typically difficult to measure complex human features (like self-confidence) with only a handful of survey items. Guryan et al. (2023) do not find evidence for effects on socioemotional skills.

5. Closing the income-achievement gap?

The HDT program aimed at improving opportunities for children growing up in low-income, disadvantaged communities in the Netherlands. We find sizable treatment effects for this group, without any clear heterogeneity within this population. The program, therefore, seems to contribute to a "neighborhood effect", boosting test scores across the board. To what extent can these programs be used to reduce the size of the achievement gap between low and high-income students?

Based on administrative data on all primary students in the Netherlands and their parents, we estimate the relationship between average standardized math achievement scores (based on the [Cito] end-of-primary-education test[31]) and parental income in Fig. 2.[32] The figure presents the familiar income-achievement gap: on average, low-income students score much lower than high-income students on cognitive achievement tests. The difference in achievement between students at the 10th percentile and the 90th percentile of the parental income distribution is 0.75 national population standard deviations.

The Dutch income-achievement gap seems only moderately less pronounced than that of the US,[33] despite differences in the level of income inequality.[34] While the support for low-income groups might contribute to reducing inequality in educational outcomes in the Netherlands, income-achievement gaps are persistent. We find that income-achievement gaps have not changed at all between 2008/09 and 2014/15, the period for which we have comparable data. Borghans et al. (2018) and Borghans and Diris (2021) draw similar conclusions based on different data and based on longer time windows.

The dashed horizontal lines in Fig. 2 indicate the average math scores for the 10th, 50th and 90th percentile of the parental income distribution. We find that the difference between the 10th and the 50th percentile is 0.27 national population standard deviations, and about equal to the one-year treatment effects we have presented in Table 4. The treatment effects of the HDT program, therefore, are relevant in

[31] The end-of-primary-education test is a different test than the tests we use as outcomes in the Tables 4–6. Both tests however are developed by test developer Cito.

[32] We have standardized test scores by subtracting the population mean and dividing by the population standard deviation after adjusting the standard deviation for less than perfect reliability. See Appendix F for derivations. Reardon (2011) applies similar adjustments. The Cito end-of-primary-education test consists of a math and literacy component. The test is widely used in the Netherlands and has been a key aspect of Dutch primary education since the 1970s. However, from 2014/15 onward other test providers could enter the market for tests. As a consequence, many schools (particularly schools with scores below the population average) now use tests from different providers. This means that comparing scores across time and between different income levels, for example, has recently become more difficult. We calculate the combined total of parental income before tax, by adding up father's and mother's income.

[33] See e.g. Reardon, 2011 and Michelmore & Dynarski, 2017.

[34] In the Netherlands, 80% of students at the 10th percentile of the parental income distribution receives rent subsidies (see Appendix G for estimates). Also, for 60% of students at the 10th percentile of the parental income distribution are so-called "weighted students". Primary schools in the Netherlands receive additional funding for weighted students (see Appendix G for estimates, see also Ladd and Fiske (2011) for a perspective on the Dutch model of primary school financing). There is also low-cost health care for all citizens and generally high-quality and freely (or cheaply) accessible primary, secondary and tertiary education throughout the Netherlands.

the context of reducing inequalities in achievement between different socioeconomic groups.

Also the targeting of low-income schools or low-income neighborhoods (with a few schools) seems practical. We have mentioned before that median parental income of students in our sample was at the 15th percentile nationally. Fig. 2 shows that mean achievement, conditional on the 15th income percentile is about 0.3 national population standard deviations below the national mean. Such neighborhoods are projected to score around the national mean on average, if HDT were to be successfully implemented there. Such low-income neighborhoods might match average achievement levels of much more affluent neighborhoods in the Netherlands.[35] Policymakers might also be interested in rolling out HDT even further, across all low-income neighborhoods in the Netherlands for example. If such an effort were to be executed successfully, our results predict that the relationship between income and achievement might be substantially flattened. Of course, this would only apply to the part of the curve to the left of the median.

6. Conclusion

We show in this paper that high-dosage math tutoring (HDT) programs can have meaningful effects on math achievement of low-income primary students in the Netherlands. We document these results against a backdrop of a persistent and sizable achievement gap between high-income and low-income students. The fact that the HDT program examined here was built from the ground up in an effort to improve opportunities in a low-income neighborhood, contributes to the idea that HDT is effective and scalable. Although the program was inspired by Saga Education in the US, and while Saga consulted on the project, the Dutch program was implemented by professionals who had no prior experience with implementing HDT. By replicating substantial effects in a different institutional context and with some adaptations to the intervention, our results add to the evidence base for HDT programs. Ongoing experimental research is studying a similar high-dosage math tutoring program in another low-income neighborhood in the Netherlands (De Ree & Paulle, 2021).

We find that the math tutoring intervention can increase math scores by an average of 0.28 national population standard deviations, enough to close the achievement gap in math between low-income and median-income primary students. We also find that high-achievers in the low-income neighborhood benefit as much as students with lower prior achievement. This suggests that high-achieving students from relatively disadvantaged backgrounds can also realize much greater academic gains than they are presently achieving.

While this research is consistent with the idea that tutoring programs are scalable and exportable, one obstacle for a further roll-out might be the non-negligible cost of implementation. Setting up an intervention with daily tutoring would cost approximately €3,000–4,000 per student per year, depending on the details of implementation. Cost issues have contributed to service providers developing alternative HDT models that are currently being implemented and evaluated. For example, Saga Education is implementing hybrid models that rely on computer aided instruction alongside professional tutors. Programs based on a half-dosage model (two or three days a week) are also currently being implemented (and RCT-tested) in the Netherlands by Stichting (Foundation) The Bridge Learning Interventions.[36] Related to this is a paper by Carlana and La Ferrara (2021) who demonstrate that the treatment effects of an online tutoring program offered during COVID-19 related lock-downs increase proportionally with the intensity of the intervention. More research is needed on the effects of reducing the dosage (or altering the delivery model) of such tutoring interventions.

Putting aside for now the possibility that costs of current and future tutoring programs may be significantly reduced (through streamlining of the program or improvements in targeting), simple cost–benefit calculations suggest that the tutoring intervention examined here might still yield net positive returns. Based on historical test score data, The Netherlands Bureau for Economic Policy Analysis (2016) estimates that a one standard deviation (combining math and reading) increase in test scores predicts a €5,000 increase in individual gross yearly earnings. With 40 years of employment, an intervention with a 0.14[37] standard deviation treatment effect might yield $40 \times 0.14 \times €5,000 = €28,000$ in additional earnings over the life cycle. With a 3% discount rate, and with 40 years of employment starting at age 24 this means €11,000 in present value terms. For the Saga Education HDT program in the US, Guryan et al. (2023) reach similar conclusions.

We see a number of areas for further research in this area. First, given the nature of our experimental design, we are not able to study the longer term effects of the program. We intend to set up new HDT experiments to measure such longer term impacts. Second, as the HDT model has an implicit socioemotional component related to the mentoring aspect of tutoring, we see opportunities for using improved measures of socioemotional (or noncognitive) development within this context. One way forward might be to use question/item level data to disentangle cognitive and noncognitive factors from the performance on a single achievement test, as Borghans and Schils (2018) have done. Moreover, further research might focus on cluster-randomized trials, where classes or schools as whole are randomly assigned to treatment (see e.g. Fryer Jr. & Howard-Noveck, 2020). The use of cluster-randomized trials would mitigate risks of control group contamination.

CRediT authorship contribution statement

Joppe de Ree: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Project administration, Supervision. **Mario A. Maggioni:** Writing – review & editing. **Bowen Paulle:** Conceptualization, Funding acquisition, Writing – review & editing, Project administration, Supervision. **Domenico Rossignoli:** Data curation, Writing – review & editing. **Nienke Ruijs:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Dawid Walentek:** Data curation, Writing – review & editing.

Data availability

Data subject to third party restrictions.

Appendix A. Baseline summary statistics

Tables 8 and 9 compare students across the T and C conditions, at baseline. The tables show that the differences between the conditions are small and generally statistically insignificant.

[35] Note that in 2013/14, the gross yearly income of parents at the 15th percentile of the parental income distribution is around €38,000. Median gross yearly income of parents is about €78,000.

[36] https://www.tbli.nl/en/home-english/

[37] i.e. a 0.28 treatment effect on math scores and a 0.00 treatment effect on reading, average out to a 0.14 unweighted effect on the combined math and reading score.

Table 8
Baseline summary statistics for the 5th grade experimental sample, pooled across cohorts.

	(1) Semester	(2) $T5$	(3) $C5$	(4) Difference	(5) p-value
Boy	1	0.48	0.52	−0.04	0.55
Boy	2	0.48	0.52	−0.04	0.55
Weighted student	1	0.34	0.35	−0.00	0.98
Weighted student	2	0.34	0.35	−0.00	0.98
Math score observed	1	0.96	0.92	0.04	0.22
Math score observed	2	0.92	0.93	−0.00	0.94
Standardized math score	1	−0.38	−0.30	−0.08	0.51
Standardized math score	2	−0.40	−0.34	−0.06	0.61
Reading score observed	1	0.96	0.91	0.04	0.15
Reading score observed	2	0.74	0.73	0.01	0.80
Standardized reading score	1	−0.47	−0.35	−0.12	0.33
Standardized reading score	2	−0.60	−0.37	−0.23	0.15
Math speed test score observed	1	0.67	0.63	0.04	0.07
Math speed test score observed	2	0.74	0.74	0.01	0.77
Raw math speed test score	1	101.38	103.03	−1.65	0.55
Raw math speed test score	2	105.87	107.85	−1.98	0.49
Survey: achievement	2	2.69	2.73	−0.04	0.63
Survey: student behavior	2	3.65	3.64	0.01	0.90
Survey: teacher-student relationship	2	3.65	3.62	0.03	0.72
Survey: parental support	2	3.56	3.55	0.00	0.97
Survey: academic development	2	3.69	3.72	−0.03	0.59
Survey: prosocial behavior	2	3.70	3.61	0.09	0.34
Survey: self-confidence	2	2.84	2.73	0.10	0.44
Survey: ambitious	2	3.88	3.73	0.15	0.16
Survey: stress	2	3.30	3.34	−0.04	0.74
Survey: organized	2	3.09	3.05	0.04	0.65

Notes. Baseline summary statistics. Column 1 indicates the semester of the school year prior to randomization. A stratum fixed effects model is used to estimate the quantities reported in the columns 2-5.

Table 9
Baseline summary statistics for the 4th grade experimental sample, pooled across cohorts.

	(1) Semester	(2) $T4$	(3) $C4$	(4) Difference	(5) p-value
Boy	1	0.64	0.48	0.16	0.05
Boy	2	0.64	0.48	0.16	0.05
Weighted student	1	0.32	0.41	−0.09	0.30
Weighted student	2	0.32	0.41	−0.09	0.30
Math score observed	1	0.98	0.90	0.08	0.00
Math score observed	2	0.99	0.92	0.07	0.01
Standardized math score	1	−0.37	−0.30	−0.07	0.73
Standardized math score	2	−0.35	−0.63	0.27	0.29
Reading score observed	1	0.98	0.92	0.06	0.02
Reading score observed	2	0.69	0.63	0.06	0.09
Standardized reading score	1	−0.39	−0.39	0.00	1.00
Standardized reading score	2	−0.45	−0.69	0.25	0.23
Math speed test score observed	1	0.38	0.35	0.03	0.28
Math speed test score observed	2	0.45	0.44	0.01	0.72
Raw math speed test score	1	105.29	96.57	8.71	0.22
Raw math speed test score	2	103.97	96.93	7.04	0.30
Survey: achievement	2	2.97	2.84	0.13	0.25
Survey: student behavior	2	3.34	3.40	−0.06	0.61
Survey: teacher-student relationship	2	3.37	3.39	−0.02	0.78
Survey: parental support	2	3.31	3.38	−0.07	0.62
Survey: academic development	2	3.55	3.58	−0.03	0.70
Survey: prosocial behavior	2	3.54	3.49	0.05	0.66
Survey: self-confidence	2	2.86	3.02	−0.17	0.30
Survey: ambitious	2	3.87	3.81	0.07	0.46
Survey: stress	2	3.12	3.19	−0.07	0.58
Survey: organized	2	3.22	3.23	−0.01	0.93

Notes. Baseline summary statistics. Column 1 indicates the semester of the school year prior to randomization. A stratum fixed effects model is used to estimate the quantities reported in the columns 2-5.

Appendix B. Robustness of control strategy

In Table 10 we present estimates of the one-year treatment effect, based on the 5th grade experimental sample. In the table, we investigate the robustness with respect to the way in which we control for baseline values. In column (1) we show results of a model without baseline controls. The estimated effects are somewhat smaller and not statistically significantly different from zero. Also, the standard errors of the column (1) results are too large to detect treatment effects in the range of 0.2–0.3 with sufficient power.

The columns (1–5) show how results change by changing some specific aspects of the control strategy. In column (2) we only use data for which first and second semester baseline outcomes are observed. In column (3) we predict missing baseline controls out of other baseline controls that are observed. For example, if the first semester baseline outcome is missing, it is predicted out of the second semester baseline outcome. In column (4) we use all observations for which we have

Table 10
Robustness of estimated one-year treatment effects on math scores for 5th grade experimental sample, with respect to changes in the control strategy.

	(1) No controls	(2) Observed controls	(3) Missing controls predicted	(4) Missing controls predicted + set to zero	(5) Control parameters flexible across samples
baseline scores of 1st and 2nd semester	0.16	0.25***	0.28***	0.24***	0.28***
	(0.12)	(0.06)	(0.06)	(0.06)	(0.06)
	[251]	[233]	[242]	[251]	[251]
baseline scores of 1st semester	0.16	0.24***	0.25***	0.22***	0.25***
	(0.12)	(0.07)	(0.07)	(0.07)	(0.07)
	[251]	[238]	[242]	[251]	[251]
baseline scores of 2st semester	0.16	0.28***	0.29***	0.25***	0.28***
	(0.12)	(0.07)	(0.07)	(0.07)	(0.07)
	[251]	[236]	[242]	[251]	[251]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Estimated treatment effects are measured in national population standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

Table 11
Treatment effects on math scores, measured in control group standard deviation units.

	(1) Half-year	(2) One year
A: Pooled 4th and 5th grade samples (3 cohorts)		
Treatment effect	0.17***	0.26***
	(0.06)	(0.07)
	[441]	[434]
B: 5th grade experimental sample (3 cohorts)		
Treatment effect	0.22***	0.25***
	(0.06)	(0.06)
	[255]	[251]
C: 4th grade experimental sample (3 cohorts)		
Treatment effect	0.10	0.28*
	(0.11)	(0.14)
	[186]	[183]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Estimated treatment effects are measured in control group standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

Table 12
Treatment effects on reading comprehension scores, measured in control group standard deviation units.

	(1) Half-year	(2) One year
A: Pooled 4th and 5th grade samples (3 cohorts)		
Treatment effect	−0.07	−0.06
	(0.07)	(0.08)
	[444]	[409]
B: 5th grade experimental sample (3 cohorts)		
Treatment effect	−0.08	−0.07
	(0.10)	(0.10)
	[256]	[222]
C: 4th grade experimental sample (3 cohorts)		
Treatment effect	−0.06	−0.05
	(0.11)	(0.11)
	[188]	[187]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Estimated treatment effects are measured in control group standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

Table 13
Treatment effects on math speed test (*Tempotoets Rekenen*), measured in control group standard deviation units.

	(1) Half-year	(2) One year
A: Pooled 4th and 5th grade samples (3 cohorts)		
Treatment effect	0.21**	0.42***
	(0.10)	(0.11)
	[424]	[419]
B: 5th grade experimental sample (3 cohorts)		
Treatment effect	0.17**	0.32***
	(0.08)	(0.09)
	[256]	[232]
C: 4th grade experimental sample (3 cohorts)		
Treatment effect	0.29	0.56**
	(0.22)	(0.22)
	[168]	[187]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Estimated treatment effects are measured in control group standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

outcomes, where the baseline data is set to zero when missing and a dummy for missing baseline data is included in the regression. In column (5) the parameter on the controls is allowed to vary across experimental samples and across cohorts. Table 10 shows that controlling for baseline values is key to a powerful design, but that the exact way of controlling for baseline values, in our view, does not make a great difference.

Appendix C. Treatment effects measured in control group standard deviation units

See Tables 11–13.

Appendix D. Heterogeneity of treatment effects by baseline achievement

See Fig. 3.

Appendix E. Testing for random attrition

Nonrandom attrition is a potential threat to internal validity. We test whether a missing test score depends on treatment assignment in Tables 14 and 15 panel A. In Tables 14 and 15 panel B we test whether a missing test score depends on the baseline test score in

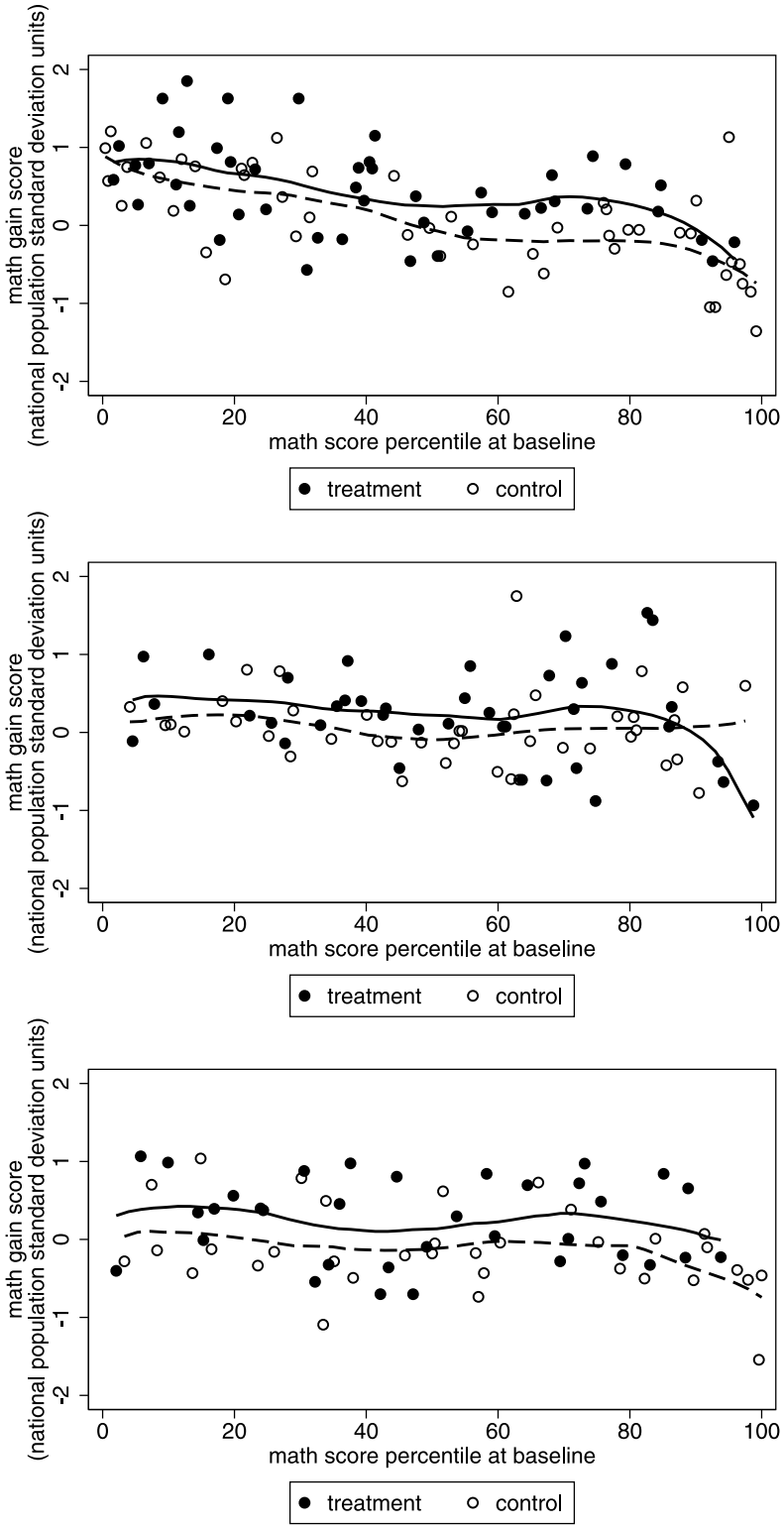

Fig. 3. Scatter diagrams of baseline to endline math gain scores, against baseline math scores in percentiles. Top, middle and bottom figures are for the 2015/16, 2016/17 and 2017/18 5th grade experimental cohorts respectively. Local linear polynomial (with bandwidth 10) are used to fit the data. The solid (dashed) curves represent the conditional mean estimates for treatment $T5$ (control $C5$).

Table 14
Tests on random attrition (predicting absence of outcome math scores). Estimates are based on the 5th grade experimental sample.

	(1) Half-year	(2) one-year
A: Simple comparison		
Treatment	−0.034	−0.019
	(0.023)	(0.026)
	[265]	[265]
B: Interacted with baseline outcome		
Treatment	−0.019	−0.001
	(0.018)	(0.022)
Baseline outcome	−0.011*	−0.019**
	(0.007)	(0.009)
Treatment × Baseline outcome	0.006	0.008
	(0.010)	(0.014)
	[252]	[252]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Student level cluster-robust standard errors in parentheses. The sample size in brackets. All regressions include stratum.

Table 15
Tests on random attrition (predicting absence of outcome reading comprehension scores). Estimates are based on the 5th grade experimental sample.

	(1)	(2)
A: Simple comparison		
Treatment	−0.011	−0.024
	(0.023)	(0.032)
	[265]	[265]
B: Interacted with baseline outcome		
Treatment	−0.006	−0.016
	(0.016)	(0.026)
Baseline outcome	−0.012	−0.026
	(0.009)	(0.016)
Treatment × Baseline outcome	−0.002	−0.007
	(0.019)	(0.027)
	[253]	[253]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. Student level cluster-robust standard errors in parentheses. The sample size in brackets. All regressions include stratum fixed effects.

the same way between treatment and control. We operationalize this by regressing an indicator for missing outcome test score data on the baseline score, an indicator for the treatment group, and an interaction between the two. We find that baseline scores are somewhat predictive of sample attrition, but we do not measure significant differences between treatment and control conditions. The tables do not provide evidence for nonrandom attrition.

Appendix F. Adjustments for less than perfect reliability

For Fig. 2 we want to estimate $E[s^*|p_k]$, where s^* is the standardized *true* math score and p_k are parental income percentiles.

Suppose that the underlying observed math score y measures the true math score y^* with random noise e:

$$y = y^* + e \tag{4}$$

with $E[e|y^*] = 0$. We show below that with these assumptions, the quantity of interest $E[s^*|p_k] = E\left[\frac{y-E[y]}{SD(y)\sqrt{\rho_y}}|p_k\right]$:

$$E[s^*|p_k] = E\left[\frac{y^* - E[y^*]}{SD(y^*)}|p_k\right] \tag{5}$$

$$= E\left[\frac{y - E[y]}{SD(y)\frac{SD(y^*)}{SD(y)}}|p_k\right] \tag{6}$$

Table 16
Heterogeneous treatment effects on math scores, by parental education. Estimates measured in national population standard deviation units.

	(1) Half-year	(2) One year
A: Pooled 4th and 5th grade samples (3 cohorts)		
Treatment effect	0.16**	0.24**
	(0.08)	(0.10)
	[441]	[434]
× weighted student	0.09	0.11
	(0.13)	(0.15)
	[441]	[434]
B: 5th grade experimental sample (3 cohorts)		
Treatment effect	0.21***	0.21***
	(0.08)	(0.08)
	[255]	[251]
× weighted student	0.11	0.19
	(0.14)	(0.14)
	[255]	[251]
C: 4th grade experimental sample (3 cohorts)		
Treatment effect	0.08	0.28
	(0.15)	(0.21)
	[186]	[183]
× weighted student	0.04	−0.01
	(0.25)	(0.30)
	[186]	[183]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. The parameters are estimated based on an extension of model (2). As additional right-hand-side variables, a dummy variable indicating a "weighted student" (a student with parents with low levels of education) and an interaction between the T_i and the dummy variable indicating a "weighted student" are included. The parameter on T_i and the parameter on the interaction between T_i and the dummy for a weighted student are reported in the table. Estimated treatment effects are measured in national population standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

$$= E\left[\frac{y - E[y]}{SD(y)\sqrt{\rho_y}}|p_k\right] \tag{7}$$

For Fig. 2 we therefore estimate $E\left[\frac{y-E[y]}{SD(y)\sqrt{\rho_y}}|p_k\right]$, where $\rho_y = \frac{V(y^*)}{V(y)}$ is the reliability of test score y. Reliability rates ρ_y are not observed in the data, but we use reliability rates reported by the test developer Cito. Generally, reliability rates for these tests are high: 0.90 for the separate math and language components and 0.95 for the full test (Cito, 2013).

Appendix G. Parental income and some dimensions of government support

Fig. 4 shows that low income students are more likely to be so-called "weighted" students (for which schools receive additional state funding) and more likely to receive rent support. The figure also shows that the lowest income percentile categories are a special group, mixing low income earners as well as others with higher earning potential, e.g. entrepreneurs with occasional low income spells. As this is beyond the scope of this paper, we have not specifically studied this group in more detail.

Appendix H. Heterogeneous treatment effects by parental education

See Tables 16 and 17.

Appendix I. Teacher questionnaire

Below, we list the questions that are used in the teacher questionnaire to measure socioemotional skills as well as concepts like

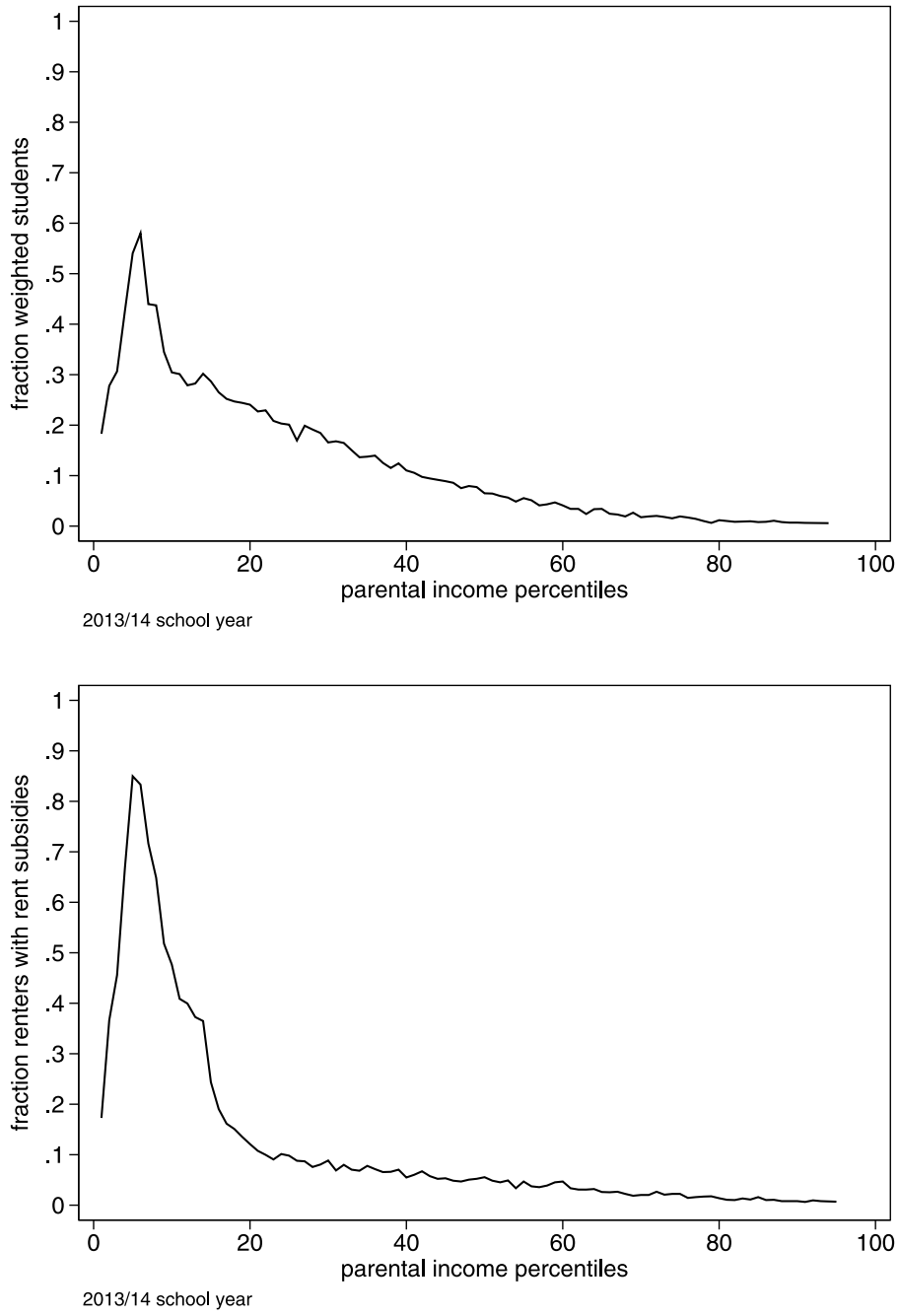

Fig. 4. Relationship between parental income of primary school 6th graders and fraction of "weighted" students [top panel] and fraction renters with rent support [bottom panel].

relationships and parental involvement. Teachers assess the statements about the students on a five point scale, ranging from "definitely untrue" to "definitely true".

Achievement

- This child is underperforming (-)
- This child can do better than he/she thinks (-)
- This child needs to work hard for their academic achievements (-)

Student behavior

- Sticks to the rules (+)
- Works accurately (+)
- Quickly thinks his/her work is done (-)
- Can get along well with classmates (+)
- Often disrupts the lessons (-)
- Has few friends in the class (-)

Teacher-student relationship

- This child is strongly focused on me (-)
- This child thinks he/she is being disadvantaged (-)
- Sometimes it takes a lot of energy to deal with this child (-)
- When this child is sad, it seeks comfort with me (+)
- There is a cultural gap between me and this child (-)
- I have a warm, caring relationship with this child (+)
- This child helps me to maintain a good classroom atmosphere (+)
- It is hard to get in touch with this child (-)

Table 17
Heterogeneous treatment effects on math speed test (*Tempotoets Rekenen*), by parental education. Estimates measured in control group standard deviation units.

	(1) Half-year	(2) one-year
A: Pooled 4th and 5th grade samples (3 cohorts)		
Treatment effect	0.25*	0.48***
	(0.13)	(0.15)
	[424]	[419]
× weighted student	−0.10	−0.19
	(0.23)	(0.25)
	[424]	[419]
B: 5th grade experimental sample (3 cohorts)		
Treatment effect	0.24**	0.36***
	(0.11)	(0.11)
	[256]	[232]
× weighted student	−0.19	−0.11
	(0.18)	(0.20)
	[256]	[232]
C: 4th grade experimental sample (3 cohorts)		
Treatment effect	0.27	0.66**
	(0.32)	(0.32)
	[168]	[187]
× weighted student	0.03	−0.30
	(0.52)	(0.53)
	[168]	[187]

Notes. ***, **, * indicate statistical significance at the 1, 5, and 10% level. The parameters are estimated based on an extension of model (2). As additional right-hand-side variables, a dummy variable indicating a "weighted student" (a student with parents with low levels of education) and an interaction between the T_i and the dummy variable indicating a "weighted student" are included. The parameter on T_i and the parameter on the interaction between T_i and the dummy for a weighted student are reported in the table. Estimated treatment effects are measured in national population standard deviation units. Student level cluster-robust standard errors in parentheses. Sample size in brackets. All regressions include stratum fixed effects and baseline outcome values. Missing baseline test scores are set to zero and a dummy variable is included in the regression model that is 1 for observations for which baseline tests are missing.

Parental involvement

- The parents are actively involved in school (+)
- There are stark differences between the home culture and the school culture (-)
- The parents support the student in learning (+)
- There are problems at home that hinder this child's progress (-)

Academic development

- You can speak Dutch with this child (+)
- There is a risk of aiming too high with this child (-)
- I am satisfied with this child's academic performance (+)
- This child can do better than he/she is right now (-)
- It pays off to put in extra effort with this child (+)
- With this child, the learning objectives should be limited (-)
- My investments in this child pay off (+)

Prosocial behavior

- Does chores immediately (+)
- Sympathizes with others (+)
- Sticks to agreements (+)
- Has little interest in others (-)
- Tries to help other people (+)

Self-confidence

- Is quiet in a group of strangers (-)
- Likes to be in the spotlight (+)
- Is the center of attention during events (+)

Ambitious

- Thinks that doing your best in school is important for your future (+)
- Wants to excel in his/her future occupation (+)
- This child wants to obtain high grades (+)

Stress

- Is easily stressed (-)
- Is easily upset (-)
- Often thinks something is going wrong or will end badly (-)

Organized

- Bursting with ideas (+)
- Likes to collect information (+)
- Uses difficult words (+)
- Does not tidy up his/her belongings (-)
- Sometimes forgets that he/she needs to do something (-)

References

Alesina, A., & Glaeser, E. L. (2005). *Fighting poverty in the US and Europe: A world of difference*. Oxford University Press.

Andrabi, T., Daniels, B., & Das, J. (2021). Human capital accumulation and disasters: Evidence from the Pakistan earthquake of 2005. *Journal of Human Resources*, 0520–10887R1.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Observer*, *13*(6), 4–16.

Borghans, L., & Diris, R. (2021). Ongelijkheid in het nederlandse onderwijs door de jaren heen. In A. Gielen, D. Webbink, & B. ter Weel (Eds.), *Preadviezen voor de koninklijke vereniging voor staathuishoudkunde*. Amsterdam: ESB & Koninklijke Vereniging voor de Staathuishoudkunde.

Borghans, L., Diris, R., & Schils, T. (2018). Sociale ongelijkheid in het onderwijs is hardnekkig. *Economisch Statistische Berichten (ESB)*, *103*(4768).

Borghans, L., & Schils, T. (2018). Decomposing achievement test scores into measures of cognitive and noncognitive skills. *Working paper*.

Carlana, M., & La Ferrara, E. (2021). Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic. *21, HKS faculty research working paper series*, (001).

Cascio, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions*. (18038), NBER.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Cito (2013). *Terugblik en resultaten 2013 eindtoets Basisonderwijs groep 8*. Arnhem: Cito.

Cito (2015). *Wetenschappelijke verantwoording van de LVS-toetsen Rekenen-Wiskunde tweede generatie: Addendum hernormering september 2013*. Arnhem: Cito.

Cook, P. J., Dodge, K., Farkas, G., Fryer Jr., R. G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H., & Steinberg, L. (2014). *The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in Chicago*. (19862), NBER.

Davis, J. M., Guryan, J., Hallberg, K., & Ludwig, J. (2017). The economics of scale-up. *NBER working paper series*, (23925).

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

De Ree, J., & Paulle, B. (2021). *High dosage math tutoring in dutch primary education: evidence from a disadvantaged neighborhood in Amsterdam*. AEA RCT Registry.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from head start. *American Economic Journal: Applied Economics*, *1*(3), 111–134.

Driessen, G., Mulder, L., Ledoux, G., Roeleveld, J., & van der Veen, I. (2009). *Cohortonderzoek COOL 5-18. Technisch rapport basisonderwijs, eerste meting 2007/2008*. Nijmegen: ITS / Amsterdam SCO-Kohnstamm Instituut.

Dutch Inspectorate of Education (2019). *De Staat van het Onderwijs 2019*. Utrecht: Inspectie van het Onderwijs.

Engzell, P., Frey, A., & Verhagen, M. (2020). Learning inequality during the COVID-19 pandemic. *Working paper*.

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

Fryer Jr., R. G., & Howard-Noveck, M. (2020). High-dosage tutoring and reading achievement: Evidence from New York City. *Journal of Labor Economics*, *38*(2), 421–452.

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., Farkas, G., Fryer Jr., R. G., Mayer, S., Pollack, H., & Steinberg, L. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, *113*(3), 738–765.

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

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

Jacob, B. A., & Lefgren, L. (2004). The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, *39*(1), 50–79.

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy*, *11*(4), 1–40.

Jungbluth, P., Roede, E., & Roeleveld, J. (2001). *Validering van het PRIMA leerlingprofiel. Reeks secundaire analyses op de PRIMA-cohort bestanden*. Amsterdam: SCO-Kohnstamm Instituut.

Kosse, F., Deckers, T., Pinger, P., Schildberg-Hörisch, H., & Falk, A. (2020). The formation of prosociality: Causal evidence on the role of social environment. *Journal of Political Economy*, *128*(2), 434–467.

Kraft, M. A. (2015). How to make additional time matter: Integrating individualized tutorials into an extended day. *Education Finance and Policy*, *10*(1), 81–116.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

Kraft, M. A., & Falken, G. T. (2021). A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*, *7*(1).

Ladd, H. F., & Fiske, E. B. (2011). Weighted student funding in the Netherlands: A model for the US? *Journal of Policy Analysis and Finance*, *30*(3), 470–498.

Michelmore, K., & Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in educational outcomes. *AERA Open*, *3*(1).

Nickow, A., Oreopoulos, P., & Quan, V. (2020). The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. *NBER working paper series*, (27476).

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. *AERA Open*, *7*(1).

Reardon, S. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In *Whither opportunity? rising inequality, schools, and children's life chances*. Routledge.

Sorrenti, G., Zölitz, U., Ribeaud, D., & Eisner, M. (2020). The causal impact of socio-emotional skills training on educational success. *IZA DP*, (13087).

The Netherlands Bureau for Economic Policy Analysis (2016). *Kansrijk onderwijsbeleid*. Den Haag: CPB.