What do teachers want? An inverse optimum approach

Thibault Deneus & Erwin Ooghe*

May 15, 2025

Abstract

We introduce a teacher time allocation model in which teachers allocate their available instruction time among individual, group, and classroom instruction to maximize a function of pupils' test scores. We consider two variants of the model, one with knowledge spillovers, the other with instruction spillovers. We evaluate both variants and find that the variant with instruction spillovers performs better, but requires more assumptions. We also derive teachers' marginal social welfare weights for their pupils and examine the influencing factors. The weights are predominantly positive, indicating teacher efficiency, decrease with higher math scores, suggesting inequality aversion, and show no significant correlation with gender, home language, or mother's education, implying anonymity. These results appear robust regardless of the presence and type of spillover effects.

Keywords: teacher preferences, marginal social welfare weights, inverse optimum, teacher time allocation, taste-based discrimination

JEL-codes: D1, D6, I2, J71

*We would like to thank Jad Beyhum, Willem De Cort and Sebastiaan Maes for helpful comments and suggestions. We also thank the audience of the 10th LEER conference for their feedback. Thibault Deneus gratefully acknowledges financial support by the Research Foundation - Flanders (FWO). E-mail addresses and affiliations: thibault.deneus@kuleuven.be (Department of Economics, KU Leuven); erwin.ooghe@kuleuven.be (Department of Economics, KU Leuven, ZEW, Mannheim, and CESifo, Munich)

1 Introduction

The distribution of pupils' educational outcomes results from the inputs of pupils, teachers, and parents, that are in turn based on their preferences and constraints. While the constraints, particularly the educational production technology, have been extensively studied, the preferences of the different agents remain largely unexplored. Given the crucial role of teachers in education, this paper aims to infer information about teachers' preferences, specifically focusing on the marginal social welfare weights assigned to their pupils.

We introduce a model of teacher time allocation in two variants. Both variants assume that teachers allocate their available instruction time to maximize a teacher-specific function of their pupils' educational outcomes. The variants differ in their approach to modeling peer effects. One variant assumes knowledge spillovers, where pupils first transform teacher instruction into knowledge and then share this knowledge with other pupils. The other variant assumes instruction spillovers, where pupils first share teacher instruction with other pupils and then transform this instruction into knowledge. We use the two variants of the model to conduct two empirical exercises.

First, we non-parametrically test the two variants. Both models perform reasonably well: the behavior of at least two-thirds of the teachers aligns with these models. The model with instruction spillovers can perform better: if we allow for peer effects, more than four-fifths of the teachers' behavior is consistent with the model. However, the model with instruction spillovers also requires a strong separability assumption that is rejected by our data.

Second, we shed light on the teachers' preferences using an inverse optimum approach. This approach allows us to infer the teachers' marginal social welfare weights for each pupil from the first-order conditions. We show that the weights (i) are mostly positive (indicating efficient teachers), (ii) decrease with test scores for most pupils (indicating inequality-averse teachers), (iii) do not depend on gender, home language, and mother's education (indicating anonymous, unbiased teachers), (iv) decrease more steeply for more experienced teachers, and (v) are relatively higher for better-performing students in the final grade.

Our paper contributes to the economics of education literature in at least two ways. To the best of our knowledge, only one other paper uses a microeconomic model of teacher time allocation to infer teachers' objectives. Moreover, our test for anonymity is, to the best of our knowledge, the first outcome test designed to detect and quantify taste-based biases of teachers in education. We discuss two related strands of the literature — one on the microeconomics of teacher preferences and the other on teacher biases in education — and highlight our contribution.

With respect to teacher preferences, there is, to the best of our knowledge, only one paper that uses a model of teacher time allocation to infer the objectives of teachers. In Brown and Saks (1987), teachers allocate individual instruction time (a private good) and classroom instruction time (a public good) to maximize the average transformed test scores of their pupils, subject to a time constraint. They then empirically investigate the effect of instruction time on math and reading scores and use these estimates to infer teacher preferences, which tend to be strongly egalitarian.

Our paper builds on their work in several ways. First, we add a third instruction mode—small group instruction—because 93% of the teachers in our data use this mode.[1] Second, we allow for corner solutions as 12.5% of the pupils in our data do not receive individual instruction time. Third, we introduce endogenous spillovers in the model. Fourth, we define the marginal social welfare weights flexibly in our model, enabling us to detect inefficient teachers (with negative weights) and non-anonymous teachers (with weights that may depend on pupil characteristics besides test scores). Fifth, we survey the teachers' marginal productivities directly, rather than estimating them from the data. This approach is closer to the idea that teachers allocate their instruction time based on their beliefs about their productivity, which may differ from actual productivity.[2] Sixth, we test the two model variants non-parametrically and compare their performance to shed light on the plausibility of the different spillover channels.

With respect to teacher biases, there is a huge literature that has proposed different ways to detect biases in different areas. Grading is one area where teacher biases could occur. Non-experimental studies compare teacher grades with central exam grades, resulting in mixed findings; see, e.g., Lindahl (2007) for Sweden, Lavy (2008) for Israel, Burgess and Greaves (2013) for England, Botelho, Madeira, and Rangel (2015) for Brazil, and Triventi (2019) and Alesina et al. (2024) for Italy. Field experiments that randomly assign pupil attributes (such as names) to exams or essays show discrimination against lower caste pupils in India (Hanna and Linden, 2012) and Turkish pupils in Germany (Sprietsma, 2013). In the Netherlands, the evidence is mixed (Van Ewijk,

[1] The more general model in Brown and Saks (1975) can also incorporate instruction in small groups, but does not bring the model to the data.

[2] It also avoids the complex issue of separately identifying teachers' preferences and constraints.

2011; Feld, Salamanca, and Hamermesh, 2015).

Besides grading, there are other areas in education where biases may occur. Black and poor students in the United States are punished more harshly than their peers involved in the same incidents (Kinsler, 2011; Barrett et al., 2021). Requests to visit a school in Spain are more likely to receive a response if the child's name is Spanish rather than Romanian (de Lafuente, 2021). Placement in special education could potentially be biased, but evidence suggests the opposite (i.e., too little placement) for minority students (National Research Council, 2002). Track placement in middle and high school may also be a source of bias with long-term consequences (see, e.g., Borghans et al., 2019; Dustmann, Puhani, and Shönberg, 2017), but evidence from the United States is mixed (Garet and Delany, 1988 versus Lucas and Gamoran, 2002).

In cases of teacher biases, whether in grading or other areas, it is important to identify who is biased against whom. One avenue of research investigates student-teacher interactions, focusing on readily observable characteristics. Dee (2005, 2007) shows that having a 'similar' teacher in terms of race or gender impacts students' achievement and engagement, and symmetrically, having a 'similar' student affects teachers' perceptions of student performance and behaviors. Ouazad and Page (2011) report that teachers in the United Kingdom tend to give better grades to students of their own gender. In contrast, Sprietsma (2013) finds no correlation between observed teacher characteristics and grading bias in Germany. Similarly, Kinsler (2011) finds little evidence that black students in the United States are punished differently based on the race of the teacher or principal. Papageorge, Gershenson, and Kang (2020) find that teachers are generally overly optimistic about their students' prospects, but white teachers are less so with black students.

Our test for anonymity can be seen as an outcome test to detect and quantify taste-based biases of teachers in education.[3] Outcome tests are popular tools to detect biases in policing and profiling (see, e.g., Persico, 2009), but have, to the best of our knowledge, not been used to detect biases in education.[4] The test is developed to detect taste-based biases that affect pupils through the preference-based choices of teachers (such as the allocation of instruction time). However, other channels cannot be detected. For example, if biases affect pupils through lowering the self-confidence or aspirations

[3]As biases are assigned to teachers' preferences, we test for taste-based discrimination (introduced by Becker, 1957) rather than statistical discrimination (introduced by Arrow, 1972 and Phelps, 1972).

[4]See Farkas (2003) for an early overview of biases in education.

of pupils, this will go undetected.[5]

The remainder of the paper proceeds as follows. In Section 2, we introduce a model of teacher time allocation along with two variants of spillovers. Section 3 presents the data, which were specifically collected for this study. Section 4 discusses the non-parametric tests of the two model variants. Section 5 derives the welfare weights and examines their determinants. A final section 6 concludes.

2 A model of teacher time allocation

We introduce two variants of a teacher time allocation model. The first variant assumes peer effects based on knowledge spillovers. For instance, an increase in the private instruction time of a pupil enhances their knowledge, which may subsequently spill over to other pupils. The second variant assumes peer effects based on instruction spillovers. In this case, an increase in the private instruction time of a pupil not only enhances their knowledge but also allows the instruction itself to spill over to other pupils, thereby improving their knowledge as well. To present both variants, we first outline their common components.

A teacher has n pupils, collected in a set $N = \{1, 2, \ldots, n\}$.[6] The set of pupils N is partitioned in m pupil groups denoted $N_1, N_2, ..., N_m$.[7] Let $k(i)$ denote the group to which pupil i belongs and let $M = \{1, 2, \ldots, m\}$ denote the set of groups.

Teachers have a total amount of instruction time T available for a given subject (e.g., math).[8] They can allocate their time to (i) individual instruction $t = (t_1, t_2, \ldots, t_n)$, with t_i the instruction that only pupil i receives (a private good), (ii) group instruction $g = (g_1, g_2, \ldots, g_m)$, with g_k the instruction that only the pupils of group k receive (a club good), (iii) classroom instruction c, the instruction that all pupils receive (a public good). We call $x_i = H_i(t_i, g_{k(i)}, c)$ the global instruction received by pupil i, with H_i a pupil-specific, differentiable, and strictly increasing function of the different

[5] See, e.g., Carlana (2019) and Papageorge, Gershenson, and Kang (2020) on the impact of teacher bias on self-confidence and the long-term impact of teacher expectations.

[6] For ease of exposition, we do not index teachers, even though all elements of the model are teacher-specific.

[7] This partitioning is assumed to be given and thus not a choice variable for the teacher. This is not entirely unrealistic, as most teachers in Flanders allow pupils to self-select in (usually three) groups depending on their need for extra teacher time (e.g., no extra time needed, possibly extra time needed, always extra time needed).

[8] The amount of hours of instruction time per subject is assumed to be exogenous to the teacher (e.g., it is fixed by, e.g., the school team, the school direction, or the school group).

instructional activities. In the remaining time $r_i = T - (t_i + g_{k(i)} + c)$, pupil i is not instructed by the teacher and processes the received individual, group and classroom autonomously. The budget constraint of the teacher is $\sum_{i \in N} t_i + \sum_{k \in M} g_k + c \leq T$.

Let s_i be the test score that pupil i achieves in the subject under consideration; the vector $s = (s_1, s_2, \ldots, s_n)$ collects all test scores. The test scores of pupils depend on the time allocation of the teacher and the peer effects. In the next two sections, we will provide more details. Each teacher allocates the available instruction time over pupils to maximize $V(s)$, a differentiable evaluation function of test scores, subject to the teacher's budget constraint and non-negativity constraints. The Lagrangian is

$$V(s) + \lambda_b(T - \sum_{i \in N} t_i - \sum_{k \in M} g_k - c) + \sum_{i \in N} \lambda_{t,i} t_i + \sum_{k \in M} \lambda_{g,k} g_k + \lambda_c c, \tag{1}$$

with $\lambda_b \geq 0$, $\lambda_{t,1}, \lambda_{t,2}, \ldots, \lambda_{t,n} \geq 0$, $\lambda_{g,1}, \lambda_{g,2}, \ldots, \lambda_{g,m} \geq 0$, and $\lambda_c \geq 0$ the multipliers of the budget constraint and the non-negativity constraints for the different instruction activities.

2.1 Knowledge spillovers

In case of knowledge spillovers, educational production is generated by[9]

$$\begin{aligned} s_i &= F_i(x_i, r_i) + P_i(s_1, s_2, \ldots, s_n), \\ &= F_i(H_i(t_i, g_{k(i)}, c), r_i) + P_i(s_1, s_2, \ldots, s_n), \end{aligned} \tag{2}$$

for each pupil i, with (i) F_i a differentiable production function of global instruction x_i and autonomous processing time r_i and (ii) P_i a differentiable peer effect function capturing knowledge spillovers.[10] As teachers allocate instruction time on the basis of beliefs, the educational production functions F_i capture what they believe they produce, not what they effectively achieve. Both are likely to converge over time, but this convergence process may proceed in different ways: teachers updating their beliefs on the basis of what their pupils achieve or pupils adjusting their realizations on the basis

[9] The production function (and, later on also the teacher's evaluation function) can also depend on the initial test scores, say, at the beginning of the school year. For ease of exposition, we do not make this dependence explicit here, but will come back to it in the empirics.

[10] Educational production is weakly separable in instructional activities and autonomous processing time. This restriction is not needed for the current model with knowledge spillovers, but is needed (and therefore already introduced here) in the model with instructional spillovers of the next section.

of the teacher's beliefs (e.g., self-fulfilling prophecies).

For ease of exposition, we abbreviate the partial derivatives of the functions V, F_i, H_i and P_i with respect to its arguments as respectively v_i, f_{ix}, f_{ir}, h_{it}, h_{ig}, h_{ic}, and p_{ij}. We assume (i) f_{ix}, f_{ir}, h_{it}, h_{ig}, $h_{ic} > 0$ and (ii) $p_{ij} \geq 0$ for all i, j, $p_{ij} = p_{ji}$ for all i, j, and $p_{ii} = 0$ for all i, and $\sum_{j \in N} p_{ij} < 1$ for all i. Moreover, we define π_{ij} as the ij-th element of the matrix $\Pi = (I - \nabla P)^{-1}$, with I the $n \times n$ identity matrix and ∇P the $n \times n$ matrix of marginal peer effects p_{ij}.[11]

Theorem 1 provides the first-order conditions of each teacher. A proof can be found in Appendix A.

Theorem 1. The first-order conditions of the Lagrangian defined in equation (1) using an educational production process with knowledge spillovers defined in equation (2) are

$$\begin{array}{rcll} (f_{jx}h_{jt} - f_{jr}) \sum_{i \in N} v_i \pi_{ij} - \lambda_b + \lambda_{t,j} & = & 0, & \text{for } j \text{ in } N, \\ \sum_{j \in N_k} (f_{jx}h_{jg} - f_{jr}) \sum_{i \in N} v_i \pi_{ij} - \lambda_b + \lambda_{g,k} & = & 0, & \text{for } k \text{ in } M, \\ \sum_{j \in N} (f_{jx}h_{jc} - f_{jr}) \sum_{i \in N} v_i \pi_{ij} - \lambda_b + \lambda_c & = & 0, & \end{array}$$

with $\lambda_b \geq 0$, $\lambda_{t,1}, \lambda_{t,2}, \ldots, \lambda_{t,n} \geq 0$, $\lambda_{g,1}, \lambda_{g,2}, \ldots, \lambda_{g,m} \geq 0$, and $\lambda_c \geq 0$.

Because group and classroom instruction are (local) public goods, we can deduce Samuelson conditions for optimal provision, requiring that the marginal rates of technical substitution (adjusted for corners solutions) must sum up to one. Corollary 1 summarizes these adjusted Samuelson conditions.

Corollary 1. The first-order conditions imply

$$\sum_{j \in N_k} \frac{f_{jx}h_{jg} - f_{jr}}{f_{jx}h_{jt} - f_{jr}} \cdot \frac{\lambda_b - \lambda_{t,j}}{\lambda_b - \lambda_{g,k}} = 1,$$

for all groups k in M and

$$\sum_{j \in N} \frac{f_{jx}h_{jc} - f_{jr}}{f_{jx}h_{jt} - f_{jr}} \cdot \frac{\lambda_b - \lambda_{t,j}}{\lambda_b - \lambda_c} = 1,$$

for the class.

[11] The assumptions on ∇P imply that the matrix Π exists, is symmetric, with non-negative elements, and its diagonal elements are strictly positive and larger than the off-diagonal elements.

2.2 Instruction spillovers

In case of instruction spillovers, educational production is generated by

$$s_i = F_i(x_i, r_i), \tag{3}$$

for each pupil i, with F_i as defined before, but instruction defined as $x_i = H_i(t_i, g_{k(i)}, c) + P_i(x_1, x_2, \ldots, x_n)$, with P_i a differentiable and non-decreasing peer effect function capturing instruction spillovers.[12] Theorem 2 provides the first-order conditions. A proof can be found in Appendix B.

Theorem 2. The first-order conditions of the Lagrangian defined in equation (1) using an educational production process with instruction spillovers defined in equation (3) are

$$\begin{aligned}
\textstyle\sum_{i\in N} v_i f_{ix} \pi_{ij} h_{jt} - v_j f_{jr} - \lambda_b + \lambda_{t,j} &= 0, \quad \text{for } j \text{ in } N,\\
\textstyle\sum_{i\in N} v_i f_{ix} \sum_{j\in N_k} \pi_{ij} h_{jg} - \sum_{i\in N_k} v_i f_{ir} - \lambda_b + \lambda_{g,k} &= 0, \quad \text{for } k \text{ in } M,\\
\textstyle\sum_{i\in N} v_i f_{ix} \sum_{j\in N} \pi_{ij} h_{jc} - \sum_{i\in N} v_i f_{ir} - \lambda_b + \lambda_c &= 0,
\end{aligned}$$

with $\lambda_b \geq 0$, $\lambda_{t,1}, \lambda_{t,2}, \ldots, \lambda_{t,n} \geq 0$, $\lambda_{g,1}, \lambda_{g,2}, \ldots, \lambda_{g,m} \geq 0$, and $\lambda_c \geq 0$.

There is no obvious way to rewrite the first-order conditions as Samuelson conditions. The reason is that one can increase, e.g., c and reduce every pupil's private time t_i to keep everyone's instruction x_i constant, but this will not necessarily keep the test scores constant as both c and x_i influence test scores via $r_i = T - t_i g_{k(i)} - c$.[13]

3 The data

To collect the data, we contacted all Flemish primary schools to participate in a survey about teacher time allocation for mathematics.[14] A total of 121 teachers from 29 schools serving more than 2500 pupils participated. The survey was conducted through a questionnaire completed by the teachers for all pupils in their classes.

[12] Note that, analogous to knowledge spillovers, instruction spillovers are frictionless, i.e., they do not come at the cost of processing time.

[13] Suppose we would model instruction spillovers as $s_i = F_i(x_i)$ with r_i with $x_i = H_i(t_i, g_{k(i)}, c, r_i) + P_i(x_1, x_2, \ldots, x_n)$ then this model would become equivalent and empirically indistinguishable to the model with knowledge spillovers.

[14] Flanders is the Dutch-speaking northern part of Belgium.

3.1 The questionnaire

The questionnaire consisted of five parts. The first part focused on the current level of mathematics of the pupils and their progress in the subject. The second part aimed to gather information on the teacher's time allocation for mathematics instruction. The third and fifth part inquired about the background of respectively pupils and teachers. In between both parts, the fourth part collected data on the class structure. The survey was conducted in Dutch; Appendix C provides an English translation of the relevant questions in each part.

In the first part, teachers were asked to evaluate (on a 6-point scale ranging from very weak to very strong) the overall mathematics knowledge and skill of their students at the beginning of the school year and currently (the survey was conducted in May, which is close to the end of the school year). Teachers were also asked to assess the current level of their pupils in mathematics as well as their ability to make progress in mathematics. For the level, they were asked to provide for each pupil a score (between 0 and 100) for mathematics if their overall mathematics knowledge and skills were tested today.[15] For the progress, teachers were asked (on a 5-point scale ranging from very slowly to very fast) how quickly each student would master a mathematics exercise in three different scenarios: (i) if explained individually to the pupil, (ii) if explained in the classroom, and (iii) if they had to study it themselves. Teachers were also asked to convert each progress level of the five-point scale (very slowly to very fast) into points (on the 0-100 scale) if they had one extra instruction hour per week.

The second part asked the teachers about how many hours per week they spend on mathematics, as well as how many minutes there are in a class hour. We also asked what percentage of their time they spend on individual instruction, classroom instruction, and self-study. Next, we asked on a five-point scale (ranging from never to always) how often they spend individual time with each student for mathematics in a typical week. Teachers were also asked to convert the items of the five-point scale into minutes.

In the third part, teachers were asked to provide the following background details for each pupil: gender, grade retention, whether the pupil speaks Dutch at home, and whether the mother has a degree in higher secondary education.[16] As teachers may not know the latter two perfectly, we introduced a four-point scale (certainly not,

[15] The survey also included questions about whether language skills form a barrier for students in learning mathematics, as well as whether there were other reasons for low performance.

[16] The latter two characteristics are also collected by the department of education (in the framework of equal educational opportunities).

	Variable	N	Mean	Std. Dev	Min	Max
Teachers						
Class info	Class size	121	21.18	4.47	11	34
	Grade	119	3.51	1.66	1	6
Time use	Weekly math hours	120	6.20	0.84	5	10
	Minutes per class hour	119	45.54	5.89	25	60
	% Individual instruction	121	34.31	18.56	0	95
	% Classroom instruction	121	45.15	18.17	0	90
	% All pupils self study	121	19.71	16.70	0	65
Experience	Experience in this grade	121	8.41	7.48	0	32
	Experience in primary school	121	13.40	9.07	0	40
	Experience in teaching	121	13.96	8.87	1	40
Demographics	Female	120	0.93	0.25	0	1
Pupils						
Pupil info	Female	2408	0.50	0.50	0	1
	Grade retention	2547	0.16	0.37	0	1
	Score	2501	75.35	19.20	0	100
	Individual time	2277	66.09	85.84	0	1400
Learning speed	Progress individual instr	2033	13.28	15.72	0	92
	Progress class instr	2043	7.91	12.06	0	90
	Progress self study	2052	4.20	10.71	-5	90

Table 1: Descriptive statistics for teachers and pupils (continuous variables)

probably not, probably yes, certainly yes). Similarly, the survey included questions in the fifth part about the following teacher characteristics: gender, education level, teaching experience, and their own background (language, education degree of mother) when they attended primary school.

In the fourth part, the survey inquired about the usual classroom setup when students work autonomously, e.g., everyone at a separate desk, pupils sit in (fixed) pairs, or in (fixed) groups. In the latter case, we also asked for group sizes.

3.2 Descriptive statistics

Tables 1 and 2 provide the descriptive statistics. In Table 1 we can see that the average class size is around 21 pupils, with the distribution of grades relatively balanced across the sample. Weekly math instruction averages around 6 hours, each hour consisting of 46 minutes of effectively teaching. The majority of instructional time is spent in classroom instruction (45.15%), followed by individual instruction time (34.31%) and self-study (19.71%).

Teacher's experience is measured in terms of years of experience in the current grade, in primary school, and in total. On average, teachers have 8 years of experience in their current grade and 13 years in primary school. The total years of teaching experience is very similar to the years in primary school. The majority of teachers in the sample are female (112 teachers), with only 8 male teachers.

For the pupils, roughly half of the sample is male (1211 pupils), and about 19.4% of the pupils have repeated a grade. The average reported math score is 75. The average time spent per pupil by the teacher (*individual time*) is 66 minutes, with substantial variation as indicated by a standard deviation of 85.84. The minimum recorded time is 0 minutes, while the maximum is 1400 minutes, suggesting large disparities in individual instruction time. Teachers in general report significantly more time spent per pupil than available for individual instruction time. In Appendix F, we discuss how we dealt with this.[17]

The study progress of pupils is captured through three distinct measures: progress in individual instruction time, progress via class instruction time, and progress in self-study. As expected, teachers report that pupils have on average the highest progress in individual instruction (*progress individual instr*), with a mean score of 13 and a maximum score of 92. Progress in class instruction (*progress class instr*) was slightly lower, averaging around 8, while progress in self-study (*progress self study*) was the lowest at 4, with a minimum of -5, indicating that some pupils experience regress in this category.

The standard deviations for these progress measures indicate substantial variability across the sample. Some teachers also seem to have misunderstood the question reporting progress of 90 points or more (with a point scale between 0 and 100). However, the theoretical model is based on ratios of these numbers. So, systematic misreporting should not be an issue if the three measures are misreported with approximately the same factor.

In Table 2, we can see that the majority of teachers hold a professional bachelor's degree (94 teachers), while only 1 teacher holds a master's degree. Most teachers spoke Dutch at home as a child (117 teachers). Regarding the mother's education, 86 teachers reported that their mother certainly had a high school diploma, while 18 reported the

[17]Note that the exact instruction time is often overstated by teachers. This plays a role only in the testing of the model for teachers who work both one-on-one and in small groups, but not for the computation of the marginal welfare weights (which are based only on whether pupils receive time) or not.

Teachers					
	High school	Prof. bachelor	Aca. bachelor	Master	
Education level	7	94	16	1	
	(5.9%)	(79.0%)	(13.4%)	(0.8%)	
	Cert. no	Prob. no	Prob. yes	Cert. yes	
Mother diploma	18	11	2	86	
	(15.4%)	(9.4%)	(1.7%)	(73.5%)	
Dutch home	1	0	1	117	
	(0.8%)	(0%)	(0.8%)	(98.3%)	
	1-on-1	Group	Both		
Individual time	8	32	78		
	(6.8%)	(27.1%)	(66.1%)		
Pupils					
	Weak	Rath. Weak	Average	Rath. Strong	Strong
Math level now	243	391	798	607	508
	(9.5%)	(15.4%)	(31.3%)	(23.8%)	(19.9%)
Math level begin year	303	427	828	529	456
	(11.9%)	(16.8%)	(32.6%)	(20.8%)	(17.9%)
	Cert. no	Prob. no	Prob. yes	Cert. yes	
Mother diploma	132	260	423	1131	
	(6.8%)	(13.4%)	(21.7%)	(58.1%)	
Dutch home	453	266	188	1334	
	(20.2%)	(11.9%)	(8.4%)	(59.5%)	
	Big issue	Small issue	No issue		
Dutch level	337	583	1623		
	(13.3%)	(22.9%)	(63.8%)		

Table 2: Descriptive statistics for teachers and pupils (categorical variables)

opposite.

On top of test scores, we also asked an ordinal question on math proficiency at the beginning and the end of the school year. In general, it seems that more pupils obtain a higher level in May then at the start of the year.

Teachers report that a majority of their pupils (59.5%) certainly speaks Dutch at home, while 20% certainly does not speak Dutch at home. For 1131 pupils the teachers report that the mother certainly holds a degree in higher secondary education, while for 132 pupils it is the opposite. In terms of language being an obstacle for learning math, 1623 pupils were reported to not have issues, for 583 it was a small issue, and for a substantial 337 pupils it was a big issue.

As is clear from these descriptive statistics, we have for certain questions some missing variables. In the sequel we will use (i) 1805 pupils with non-missing variables on all relevant questions to test the model and (ii) 1189 pupils with computable marginal welfare weights to analyze its drivers. The descriptive statistics for these pupils are reported in Appendices D.1 and D.2, respectively.

4 Testing the model

The data reveals that classroom instruction time is strictly positive for all teachers (so, $c > 0$ and hence $\lambda_c = 0$). Moreover, the data also reveals that some pupils or groups of pupils get private or group instruction time (so, $t_j > 0$ and hence $\lambda_{t,j} = 0$ or $g_k > 0$ and hence $\lambda_{g,k} = 0$). Also the budget constraint holds with equality (hence $\lambda_b > 0$).[18]

Let $[\cdot]$ be equal to one if the statement between brackets is true and zero otherwise. Corollary 3 rewrites the first-order conditions of theorems 1 and 2.

Corollary 2. The first-order conditions of theorem 1 can be written as (corollary 2.1)

$$\begin{array}{rcll}(f_{jx}h_{jt} - f_{jr})\sum_{i\in N}\tilde{v}_i\pi_{ij} - 1 + \tilde{\lambda}_{t,j}\times[t_j = 0] & = & 0, & \text{for } j \text{ in } N,\\ \sum_{j\in N_k}(f_{jx}h_{jg} - f_{jr})\sum_{i\in N}\tilde{v}_i\pi_{ij} - 1 + \tilde{\lambda}_{g,k}\times[g_k = 0] & = & 0, & \text{for } k \text{ in } M,\\ \sum_{j\in N}(f_{jx}h_{jc} - f_{jr})\sum_{i\in N}\tilde{v}_i\pi_{ij} - 1 & = & 0, & \end{array}$$

[18]The budget constraint holds either by assumption or by construction. Appendix F provides more details on the time assignment in the data.

and the first-order conditions of theorem 2 can be written as (corollary 2.2)

$$\begin{array}{rcll} \sum_{i\in N}\tilde{v}_i f_{ix}\pi_{ij}h_{jt}-\tilde{v}_j f_{jr}-1+\tilde{\lambda}_{t,j}\times[t_j=0] & = & 0, & \text{for } j \text{ in } N,\\ \sum_{i\in N}\tilde{v}_i f_{ix}\sum_{j\in N_k}\pi_{ij}h_{jg}-\sum_{i\in N_k}\tilde{v}_i f_{ir}-1+\tilde{\lambda}_{g,k}\times[g_k=0] & = & 0, & \text{for } k \text{ in } M,\\ \sum_{i\in N}\tilde{v}_i f_{ix}\sum_{j\in N}\pi_{ij}h_{jc}-\sum_{i\in N}\tilde{v}_i f_{ir}-1 & = & 0, & \end{array}$$

with $\tilde{\lambda}_{t,j}=\lambda_{t,j}/\lambda_b\geq 0$, $\tilde{\lambda}_{g,k}=\lambda_{g,k}/\lambda_b\geq 0$, and $\tilde{v}_i=v_i/\lambda_b$.

Before we discuss the results, we add some remarks.

First, besides data on f_{jr}, we collected data on the products $f_{jx}h_{jt}$, $f_{jx}h_{jg}$, $f_{jx}h_{jc}$ for the different pupils of each teacher. While these products suffice to test the model with knowledge spillovers (corollary 2.1), we need the separate factors to test the model with instruction spillovers (corollary 2.2). To do so, we assume $h_{jt}=1$ such that the reported products $f_{jx}h_{jt}$ allow to deduce f_{jx}. We then compute the h_{jc}'s by dividing the reported $f_{jx}h_{jc}$ by the deduced f_{jx}. Finally, as we do not collect data on the productivity of instruction in small groups, we assume that it is more efficient than classroom instruction, but less efficient than individual instruction, that is $h_{jg}=h_{jc}+\alpha(1-h_{jc})$ with α (an unknown parameter) between zero and one.

Second, to test the two models in their most flexible way we plug in the known variables (f_{jx}, f_{jr}, h_{jt}, h_{jg}, h_{jc}, t_j, g_k) in the first-order conditions and check whether there exist unknown variables ($\tilde{v}_i$, π_{ij}, $\tilde{\lambda}_{t,j}$, $\tilde{\lambda}_{g,k}$, and α) such that the first-order conditions are satisfied. If we replace $\tilde{v}_i\pi_{ij}$ by δ_j everywhere in corollary 2.1, then it is clear that testing the first-order conditions does not require to search for unknown variables $\tilde{v}_i$ and π_{ij}, but for unknown δ_j. This considerably simplifies the test, but also highlights that the underlying welfare weights and peer effects do not matter to test the model with knowledge spillovers. This is not true, however, for the model with instruction spillovers, so we will test it under different peer effect restrictions. The most restrictive version is to assume that there are no peer effects and the least restrictive version is to assume that the peer effects can be any non-negative scalar (flexible peer effects). In between, we consider two possibilities. One possibility is called fixed peer effects: it assumes that the off-diagonal elements of the matrix ∇P are the same.[19] Another possibility is called block peer effects: it assumes that only pupils who receive the same individual instruction time (an ordinal variable) influence each other as they are more likely to sit together. So, the different blocks of pupils will have the same peer effect,

[19] This additional property implies that the diagonal elements of the matrix Π are the same and that the off-diagonal elements are the same.

knowledge spillovers	1–on–1	group	both	overall
all peer effects	0.0%	80.77%	67.65%	67.68%
instruction spillovers	1-on-1	group	both	overall
no peer effects	0.0%	80.77%	67.65%	67.68%
block peer effects	60.00%	92.31%	73.53%	77.78%
fixed peer effects	20.0%	92.31%	79.41%	79.80%
flexible peer effects*	60.00%	92.31%	79.41%	81.82%

*The reported percentages can best be interpreted as lower bounds.

Table 3: Percentages of teachers whose behavior is consistent with the different models

which may differ over the blocks. These two intermediate possibilities cannot be ranked a priori: block peer effects are more flexible in one dimension (peer effects may differ over blocks), but less flexible in another (no peer effects between blocks).

Third, while all teachers allocate time to classroom instruction, not all teachers allocate time to individual pupils (1–to–1) or to small groups of pupils (groups). Out of 99 teachers, 68 teachers use both methods. Among the remaining 31 teachers, 5 teachers mostly employ 1-to-1 and 26 teachers mostly use groups.[20]

Table 3 shows the percentage of teachers that satisfies the first-order conditions under different peer effect models (in rows). Testing the model with instruction spillovers and (flexible) peer effects is numerically challenging, so the reported percentages can best be interpreted as lower bounds.[21]

First, we focus on the overall performance reported in the last column. Recall that the peer effect structure does not matter for the model with knowledge spillovers. We find that two thirds (67.68%) of the teachers can satisfy the first-order conditions of the model with knowledge spillovers (irrespective of the peer effect structure). Because both models coincide if there are no peer effects, this percentage exactly returns if we test the model with instruction spillovers with no peer effects. Yet, if we allow for peer effects, this percentage further increases. The intermediate fixed and block peer effect structures have only one extra degree of freedom, but add at least 10% of

[20]For the 31 teachers who do not instruct both individual pupils and small groups of pupils, the corresponding time variables are set to zero in the model.

[21]For fixed and block peer effects we know enough properties of the inverse of the peer-effect matrix that we can estimate this directly. For the free peer- effect matrix we have not yet found a way to do this. Therefore, we need to invert the peer effect matrix in optimization that yields a poorly conditioned objective function.

teachers that now also satisfy the conditions. Fixed peer effects seem to offer slightly more flexibility than block peer effects (2% extra teachers). Fully flexible peer effects require many extra degrees of freedom, but only seem to offer a marginal advantage over block and fixed peer effects (another 2% extra teachers compared to fixed peer effects). Overall, the model with instruction spillovers fares better, but, as explained before, it also assumes a weakly separable structure of the educational production that is not required in case of knowledge spillovers. So, its out-performance is 'bought' by imposing functional form restrictions.

Second, we look at the performance for the different types of private instruction reported in the first three columns. Both models have difficulties to justify the behavior of (the limited number of) teachers who use a one-to-one method. The opposite is true for the teachers who mostly use small groups.

5 The marginal welfare weights and its drivers

In this section we compute and analyze the marginal welfare weights of teachers (normalized by the Lagrange multiplier, that is, v_i/λ_b's) based only on the first order conditions for individual instruction time. We split up our analysis into two parts, depending on whether we include or not peer effects. In each part, we will investigate whether teachers are efficient (positive weights), inequality-averse (weights that decrease with test scores), and impartial (same weights for pupils with the same test score). In the final part, we examine whether there is substantial heterogeneity in these welfare weights across grade levels and teacher experience.

5.1 Without peer effects

In the absence of peer effects, both models coincide. We first compute the marginal welfare weights (for pupils whose individual instruction time is non-zero). Figure 1 shows the histogram. The percentage of pupils with strictly positive weights is equal to 98%. Hence, for 98% of the pupils, teachers satisfy the monotonicity principle that higher scores are better.

Second, we investigate how the weights vary over math scores using a flexible spline, i.e., a function defined by polynomials that are estimated over intervals of math scores.[22]

[22]We also include teacher fixed effects, so the exact econometric specification is $w_{ij} = \alpha_j + f(s) + \varepsilon_{ij}$, with $w_{ij} = v_{ij}/\lambda_{bj}$ the weight of pupil i of teacher j, α_j the teacher-fixed effects, f the spline, and ϵ_{ij}

Figure 1: Histogram of the marginal welfare weights

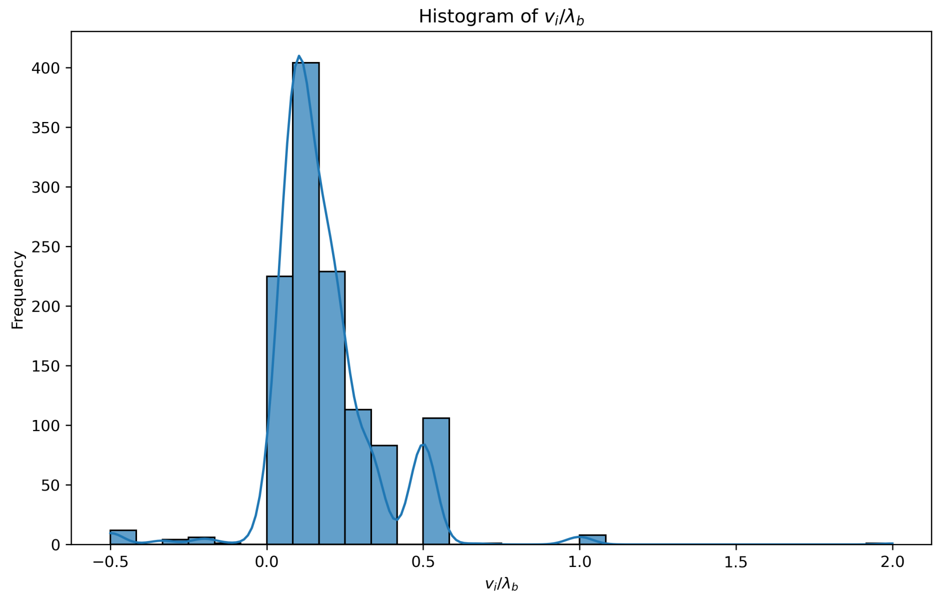


Figure 2 shows nine splines based on polynomials of second, third or fourth degree estimated over two, three, or four math score intervals (based on quantiles). The dashed vertical lines indicate the quintiles of the math score distribution. We focus on the middle panel (third degree polynomials estimated over three intervals).[23] The weights tend to first increase for pupils with very low scores (3.78% of the pupils between 0 and 30) and to decrease again afterwards. The decrease is initially quick (for 31.60% of the pupils between 30 and 70), then flat or even slightly increasing (for 50.88% of the pupils between 70 and 90), and then quickly decreasing again (for 13.75% of the pupils above 90). While standard errors are large, the spline suggests that teachers are inequality averse over a large interval (for 96.22% of the pupils above 30), but only in a very mild way as for most (60 % to 80%) pupils the spline is rather flat. In other words, they give priority to pupils with lower math scores, ceteris paribus.

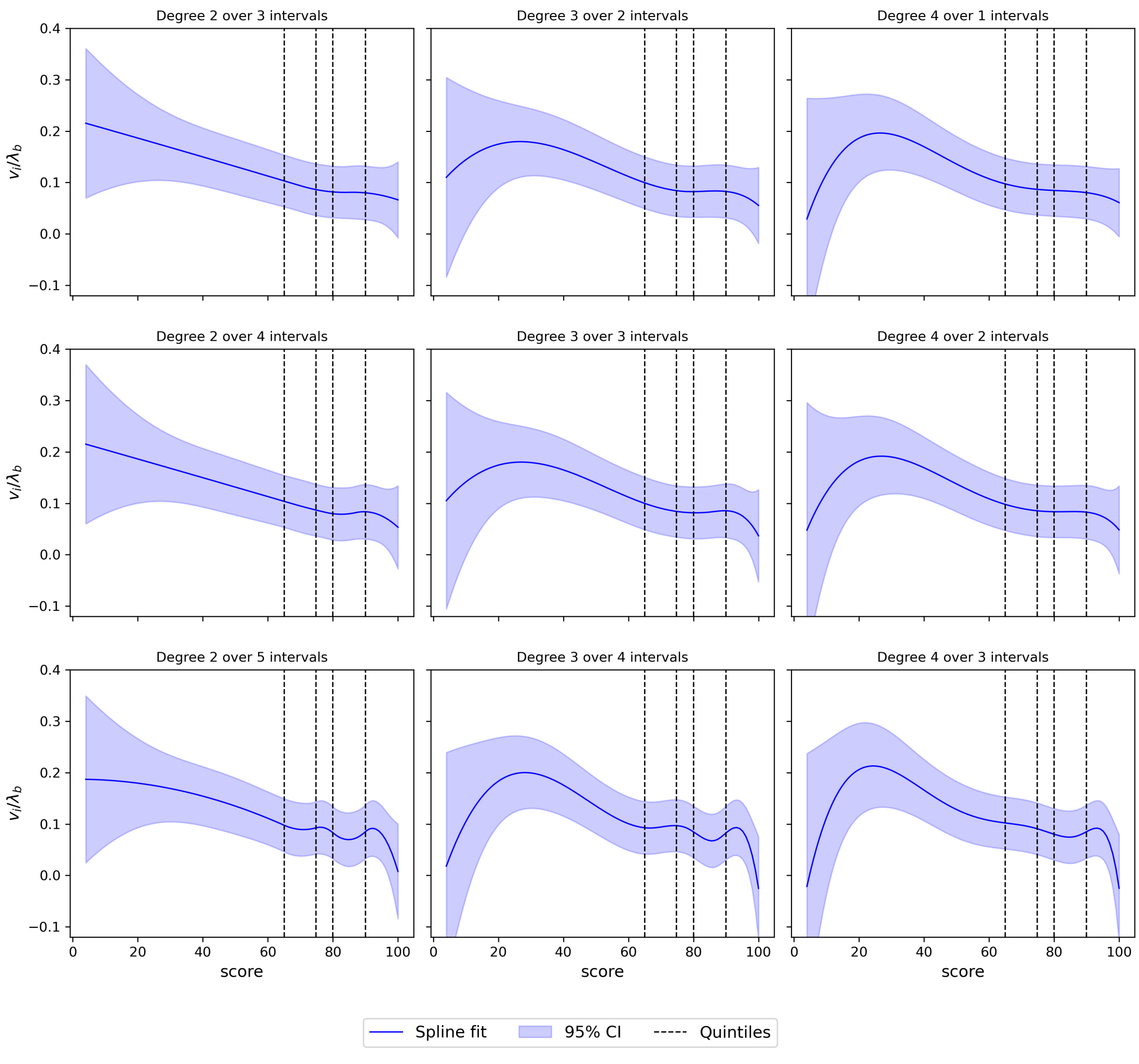
Third, we include other pupil variables in addition to the spline for math scores.[24] Table 4 shows the results. First, the impact of initial math levels (dummies based on a five-point scale) is negative and statistically significant. A negative sign means

error terms.

[23]The pattern that we describe occurs in the middle and right-hand panels (based on third and fourth degree polynomials), but not in the left-hand panels (based on second-degree polynomials) where we observe a steady decrease over the whole interval.

[24]We again include teacher fixed effects and use the middle spline of Figure 2 (with third degree polynomials estimated over three intervals) as the benchmark spline.

Figure 2: Marginal welfare weights as a function of math scores (without peer effects)



that, among two pupils with the same math level at the end of the year, teachers give a higher priority to the pupil that was initially weaker. If anything, we expected a positive sign, which would reflect that teachers do not only care about current math levels, but also about their progress. Indeed, among two pupils with the same math level at the end of the year, teachers would then give a higher priority to the pupil that was initially stronger (and hence that made less progress). Second, none of the other pupil variables (gender, Dutch at home, degree mother) are statistically significant, suggesting that there is no taste-based discrimination on the basis of these variables. Third, the inclusion of the spline does not seem to affect the results significantly.[25]

5.2 With peer effects

To deal with peer effects, we proceed as follows. We first resample with replacement (bootstrap) our data 200 times at the class level (keeping the total number of classes constant). For each bootstrap sample, we use a grid for the peer effect parameters leading to either 100 or 1028 possible peer effect matrices (Π's) in case of, respectively, fixed and block peer effects.[26] We then compute, for each matrix, the teachers' welfare weights for their pupils. For each resulting vector of welfare weights we can repeat the previous exercise, that is, estimate a spline (to visualize how the weights vary over math scores), with or without other pupil characteristics (to test for taste-based biases of teachers).

Each spline in Figure 3 is a third degree polynomial estimated over three intervals (the benchmark case that we also discussed in the previous section). The panels to the left are based on knowledge spillovers and the panels to the right on information spillovers. The patterns are very similar, but the standard errors are somewhat lower in case of information spillovers. Next, the top panels are based on fixed peer effects and the bottom panels on block peer effects. Again the differences are small: block peer effects seem to flatten the pattern somewhat, especially for high math scores, but add more noise. Finally, if we compare the splines of Figure 3 with the middle one in Figure 2 that uses the same specification, the differences are negligible.

Tables 5 and 6 present (bootstrapped) regression results under fixed and block knowledge spillovers, respectively. Each table shows the mean of the estimated coef-

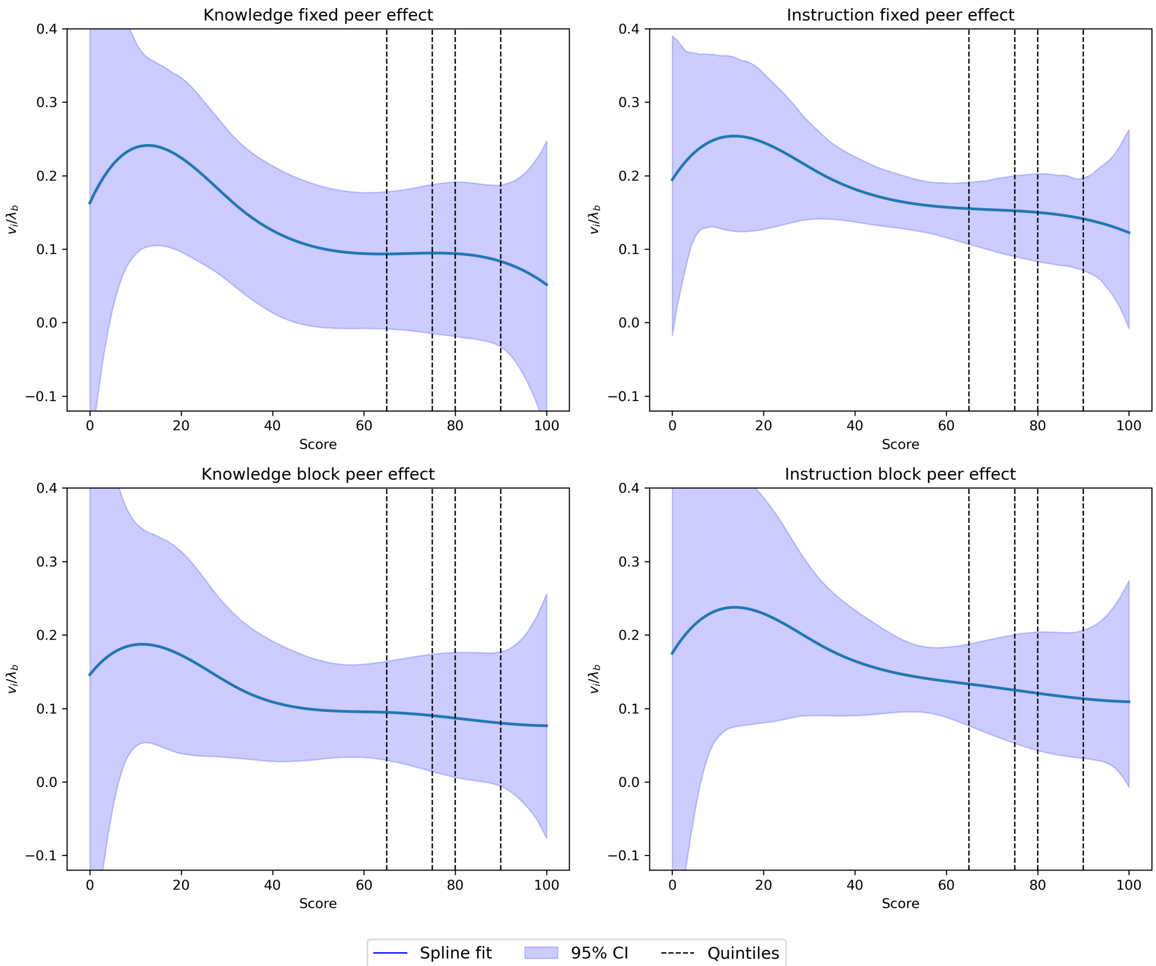
[25] Of course, the dummies for initial math level become somewhat stronger as expected.

[26] For fixed peer effects, there is one peer effect parameter, between 0 and 1, leading to a grid $0, 0.01, \ldots, 0.99$ of 100 values. For block peer effects, there are 4 parameter values in each of the 5 blocks leading to 1024 combinations.

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5
Math level begin year					
Rather weak	-0.027*	-0.027*	-0.027*	-0.029**	-0.043***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.013)
Average	-0.043***	-0.043***	-0.043***	-0.050***	-0.064***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.012)
Rather Strong	-0.046**	-0.046**	-0.046**	-0.053***	-0.066***
	(0.018)	(0.019)	(0.019)	(0.019)	(0.014)
Strong	-0.046**	-0.046**	-0.047**	-0.056**	-0.067***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.015)
Female		-0.001	-0.001	-0.000	-0.001
		(0.007)	(0.007)	(0.007)	(0.007)
Dutch home					
Probably no			-0.012	-0.015	-0.015
			(0.015)	(0.015)	(0.015)
Probably yes			-0.000	0.002	0.000
			(0.019)	(0.019)	(0.019)
Certainly yes			-0.004	-0.007	-0.007
			(0.012)	(0.013)	(0.013)
Mother diploma					
Probably no				0.026	0.022
				(0.018)	(0.018)
Probably yes				-0.019	-0.021
				(0.018)	(0.018)
Certainly yes				0.020	0.015
				(0.017)	(0.017)
Spline	Yes	Yes	Yes	Yes	No
Teacher Fixed Effects	Yes	Yes	Yes	Yes	Yes

Table 4: Drivers of the marginal welfare weights (without peer effects)

Figure 3: Marginal welfare weights as a function of math scores (with peer effects)



ficients together with the 95% confidence intervals. Regression results for (fixed and block) instruction peer effects turn out to be very similar and can be found in Appendix G.

Table 5 shows the results for fixed knowledge spillovers. Compared to the regressions without peer effects, no new insights are obtained. The estimated coefficients for the initial math scores are somewhat larger (in absolute value), but also the standard errors are somewhat larger. Also the sign and magnitude of the estimated coefficients for the other pupil characteristics remain similar, suggesting no taste-based biases. Omitting the spline leads to a strong impact of initial math level, stronger than without peer effects.

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5
Math level begin year					
Rather weak	-0.034	-0.033	**-0.035**	**-0.037**	**-0.055**
	(-0.077, 0.001)	(-0.074, 0.005)	(-0.077, -0.001)	(-0.080, -0.001)	(-0.098, -0.023)
Average	**-0.055**	**-0.055**	**-0.053**	**-0.064**	**-0.084**
	(-0.112, -0.003)	(-0.111, -0.005)	(-0.108, -0.008)	(-0.132, -0.016)	(-0.143, -0.029)
Rather Strong	**-0.056**	**-0.060**	**-0.057**	**-0.070**	**-0.086**
	(-0.118, -0.005)	(-0.127, -0.004)	(-0.119, -0.004)	(-0.152, -0.015)	(-0.161, -0.031)
Strong	-0.053	**-0.057**	-0.056	**-0.071**	**-0.090**
	(-0.125, 0.018)	(-0.121, -0.000)	(-0.120, 0.024)	(-0.151, -0.000)	(-0.170, -0.011)
Female		-0.002	-0.000	-0.001	-0.000
		(-0.025, 0.019)	(-0.022, 0.021)	(-0.023, 0.020)	(-0.024, 0.018)
Dutch home					
Probably no			-0.016	-0.019	-0.018
			(-0.038, 0.004)	(-0.041, 0.002)	(-0.045, 0.004)
Probably yes			-0.001	0.003	0.003
			(-0.028, 0.025)	(-0.027, 0.038)	(-0.031, 0.032)
Certainly yes			-0.006	-0.009	-0.008
			(-0.030, 0.019)	(-0.032, 0.017)	(-0.030, 0.013)
Mother diploma					
Probably no				0.031	0.027
				(-0.012, 0.082)	(-0.010, 0.077)
Probably yes				-0.024	-0.025
				(-0.098, 0.038)	(-0.087, 0.034)
Certainly yes				0.025	0.018
				(-0.024, 0.085)	(-0.024, 0.073)
Spline	Yes	Yes	Yes	Yes	No
Teacher Fixed Effects	Yes	Yes	Yes	Yes	Yes

Table 5: Drivers of the marginal welfare weights (with fixed knowledge peer effects).

Table 6 shows the results for block knowledge spillovers. Compared to the other regression results (without peer effect and with fixed knowledge peer effects), the estimates for initial math level are lower (in absolute value) and no longer significant. Estimates for the remaining variables remain largely stable however. None of the coefficients are significant at the 95% level, still, initial level continues to appear most informative. Omitting the spline leads again to a stronger impact of initial math level, but, as mentioned before, none of the estimates are statistically significant.

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5
Math level begin year					
Rather weak	-0.013	-0.013	-0.014	-0.016	-0.027
	(-0.050, 0.025)	(-0.051, 0.026)	(-0.052, 0.021)	(-0.056, 0.021)	(-0.073, 0.019)
Average	-0.029	-0.029	-0.029	-0.039	-0.047
	(-0.090, 0.033)	(-0.090, 0.033)	(-0.088, 0.031)	(-0.102, 0.024)	(-0.127, 0.035)
Rather Strong	-0.033	-0.036	-0.034	-0.045	-0.050
	(-0.115, 0.049)	(-0.118, 0.048)	(-0.115, 0.050)	(-0.130, 0.038)	(-0.155, 0.056)
Strong	-0.038	-0.041	-0.041	-0.053	-0.061
	(-0.140, 0.062)	(-0.137, 0.057)	(-0.141, 0.056)	(-0.157, 0.046)	(-0.176, 0.054)
Female		-0.001	0.000	-0.001	0.001
		(-0.019, 0.017)	(-0.017, 0.019)	(-0.018, 0.017)	(-0.017, 0.017)
Dutch home					
Probably no			-0.013	-0.016	-0.016
			(-0.041, 0.016)	(-0.044, 0.011)	(-0.046, 0.013)
Probably yes			-0.001	0.001	0.002
			(-0.038, 0.035)	(-0.038, 0.040)	(-0.039, 0.040)
Certainly yes			-0.008	-0.014	-0.012
			(-0.028, 0.015)	(-0.037, 0.010)	(-0.036, 0.009)
Mother diploma					
Probably no				0.021	0.019
				(-0.017, 0.060)	(-0.016, 0.056)
Probably yes				-0.023	-0.023
				(-0.088, 0.029)	(-0.082, 0.028)
Certainly yes				0.022	0.016
				(-0.018, 0.066)	(-0.022, 0.058)
Spline	Yes	Yes	Yes	Yes	No
Teacher Fixed Effects	Yes	Yes	Yes	Yes	Yes

Table 6: Drivers of the marginal welfare weights (with block knowledge peer effects).

Overall, introducing peer effects seems to add noise, as expected, but does not change the main results.

5.3 Heterogeneity in welfare weights

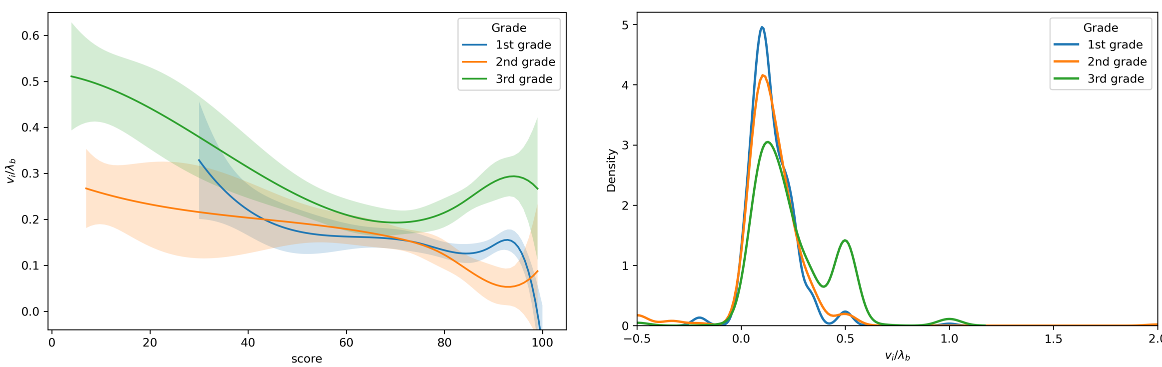
In this section, we examine whether there is heterogeneity in welfare weights by the grade level of the class and the teacher's overall teaching experience. We abstract from peer effects in this analysis because, based on the previous section, we believe their inclusion would only introduce noise. For the grade-level analysis, we group the first and second years of primary school together as first grade, the third and fourth years as second grade, and the fifth and sixth years together as third grade.

The results by grade level are presented in Figure 4. As shown in Figure 4a, the third grade has substantially higher welfare weights at both the lower and upper ends of the test score distribution. The weights for third grade also appear to increase at the right tail, suggesting that teachers place greater value on high performance in the final grade. In contrast, the welfare weights for first and second grades are mostly monotonically decreasing across the score distribution. The slight increase in welfare weights at the lower end of the score distribution observed in earlier sections appears to have disappeared. Figure 4b displays the kernel density estimates of the welfare weights. The higher average weights in third grade appear to be driven by a second peak around 0.5, which is less prominent in the other groups. Additionally, the second grade appears to have the highest number of violations of the Pareto principle.

Figure 4: Heterogeneity per grade level

(a) $\frac{v_i}{\lambda_b}$ as a function of test scores

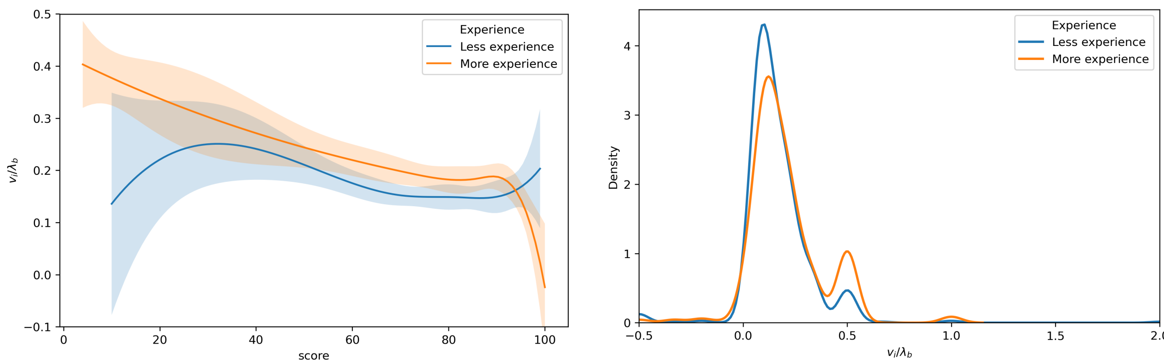
(b) Kernel density of $\frac{v_i}{\lambda_b}$



For the heterogeneity analysis by teaching experience, we divide teachers into two groups: those with more and those with less than the median level of experience. The results are presented in Figure 5. As shown in Figure 5a, more experienced teachers assign higher welfare weights to their students, except for the highest-performing ones. Their welfare weights decrease monotonically across the score distribution, suggesting greater inequality aversion. In contrast, less experienced teachers appear less inequality-averse, with lower weights assigned to students at the bottom of the score distribution. Figure 5b shows that less experienced teachers violate the Pareto principle more frequently. Meanwhile, more experienced teachers exhibit a more profound second peak around 0.5 in the distribution of welfare weights.

Figure 5: Heterogeneity per teacher experience

(a) $\frac{v_i}{\lambda_b}$ as a function of test scores (b) Kernel density of $\frac{v_i}{\lambda_b}$



6 Conclusion

We introduced a teacher time allocation model in which teachers allocate their available instruction time over individual, group, and classroom instruction to maximize a function of pupils' test scores. We combine this model with two different views on peer effects based on either knowledge or instruction spillovers.

We collect data on the time allocation and the marginal productivities of the different instruction modes of Flemish teachers in primary education to test the optimality conditions under different peer effect structures. The model with instruction spillovers performs better overall, but also requires more assumptions.

We also infer the teachers' marginal social welfare weights of their pupils in both model variants and analyze the drivers. In the absence of peer effects, the weights are (almost always) strictly positive (hence, teachers are efficient), decrease with math scores for most pupils (teachers are inequality averse), and do not significantly depend on some other pupil variables (teachers are impartial with respect to gender, home language, and mother's education). Peer effects do not seem to alter these results. We find that teachers in the final grade place greater weight on high-achieving students compared to those in earlier grades. Additionally, more experienced teachers appear to be more inequality-averse, assigning higher welfare weights to lower-performing students.

References

Alesina, A., Carlana, M., La Ferrara, E., Pinotti, P., 2024, Revealing stereotypes: evidence from immigrants in schools, *American Economic Review*, 114(7), 1916–48.

Arrow, K., 1972, Some mathematical models of race discrimination in the labor market, in: Pascal, A., ed., *Racial Discrimination in Economic Life*, Lexington: Lexington Books.

Barrett, N., McEachin, A., Mills, J., Valant, J., 2021, Disparities and discrimination in student discipline by race and family income, *Journal of Human Resources*, 56(3), 711-748.

Bertrand, M., Chugh, D., Mullainathan, S., 2005, Implicit discrimination, *American Economic Review*, 95(2), 94-98.

Becker, G., 1957, *The Economics of Discrimination*, Chicago: Chicago University Press.

Borghans, L., Diris, R., Smits, W., de Vries, J., 2019, The long-run effects of secondary school track assignment, *PLoS ONE*, 14(10), e0215493.

Botelho, F., Madeira, R., Rangel, M., 2015, Racial discrimination in grading: evidence from Brazil, *American Economic Journal: applied economics*, 7(4), 37-52.

Brown, B., Saks, D., 1975, The production and distribution of cognitive skills within schools, *Journal of Political Economy*, 83(3), 571-593.

Brown, B., Saks, D., 1987, The microeconomics of the allocation of teachers' time and student learning, *Economics of Education Review*, 6(4), 319-332.

Burgess, S., Greaves, E., 2013, Test scores, subjective assessment, and stereotyping of ethnic minorities, *Journal of Labor Economics*, 31(3), 535-576.

Carlana, M., 2019, Implicit stereotypes: evidence from teachers' gender bias, *Quarterly Journal of Economics*, 134(3), 1163-1224.

Dee, T., 2005, A teacher like me: does race, ethnicity, or gender matter? *American Economic Review, Papers and Proceedings*, 95(2), 158-165.

Dee, T., 2007, Teachers and the gender gaps in student achievement, *Journal of Human Resources*, 42(3), 528-554.

de Lafuente, D., 2021, Cultural assimilation and ethnic discrimination : an audit study with schools, *Labour Economics*, 72, 1-19.

Dustmann, C., Puhani, P., Shönberg, U., 2017, The long-term effects of early track choice, *Economic Journal*, 127, 1348-1380.

Farkas, G., 2003, Racial disparities and discrimination in education: what do we know, how do we know it, and what do we need to know? *Teacher College Record*, 105(6), 1119-1146.

Feld, J., Salamanca, N., Hamermesh, D., 2015, Endophilia or exophobia: beyond discrimination, *Economic Journal*, 126, 1503-1527.

Garet, M., DeLany, B., 1988, Students, courses, and stratification, *Sociology of Education*, 61, 61-77.

Hanna, R., Linden, L., 2012, Discrimination in grading, *American Economic Journal: Economic Policy*, 4(4), 146-168.

Kinsler, J., 2011, Understanding the black-white school discipline gap, *Economics of Education Review*, 30, 1370-1383.

Lavy, V., 2008, Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural field experiment, *Journal of Public Economics*, 92, 2083-2105.

Lindahl, E., 2007, Comparing teachers' assessments and national test results: evidence from Sweden, Working paper 2007:24, Institute for Labour Market Policy Evaluation, Uppsala.

Lucas, S., Gamoran, A., 2002, Track assignment and the achievement gap, in: Chubb, J., Loveless, T., eds., *Bridging the Achievement Gap*. Washington, DC: The Brookings Institute.

National Research Council, 2002, *Minority Students in Special and Gifted Education*, Washington, D.C.: The National Academies Press.

Ouazad, A., Page, L., 2011, Estimating perceptions of discrimination: experimental economics in schools, working paper.

Paluck, E., Green, D., 2009, Prejudice reduction: what works? A review and assessment of research and practice, *Annual Review of Psychology*, 60, 339-367.

Papageorge, N., Gershenson, S., Kang, K., 2020, Teacher expectations matter, *Review of Economics and Statistics*, 102(2), 234-251.

Papay, J., Murnane, R., Willett, J., 2016, The impact of test score labels on human-capital investment decisions, *Journal of Human Resources*, 51(2), 357-388.

Persico, N., 2009, Racial profiling? Detecting bias using statistical evidence, *Annual*

Review of Economics, 1, 229-253.

Phelps, E., 1972, The statistical theory of racism and sexism, *American Economic Review*, 62, 659-661.

Sprietsma, M., 2013, Discrimination in grading: experimental evidence from primary school teachers, *Empirical Economics*, 45, 523-538.

Triventi, M., 2019, Are children of immigrants graded less generously by their teachers than natives, and why? Evidence from student population data in Italy, *International Migration Review*, 54(3), 765-795.

Van Ewijk, R., 2011, Same work, lower grade? Student ethnicity and teachers' subjective assessments, *Economics of Education Review*, 30, 1045-1058.

A Proof of theorem 1

In case of knowledge spillovers, the test scores are defined by

$$s_i = F_i(x_i, r_i) + P_i(s_1, s_2, \ldots, s_n),$$

with $x_i = H_i(t_i, g_{k(i)}, c)$ and $r_i = T - t_i - g_{k(i)} - c$. Let $S_i(t, g, c)$ denote the test score solution for pupil i. We thus have

$$S_i(t, g, c) = F_i(H_i(t_i, g_{k(i)}, c), T - t_i - g_{k(i)} - c) + P_i(S_1(t, g, c), \ldots, S_n(t, g, c)),$$

for all i in N.

First, for private instruction time, we have

$$\frac{\partial S_i(t,g,c)}{\partial t_j} = 1[i = j](f_{ix}h_{it} - f_{ir}) + \sum_{\ell \in N} p_{i\ell} \frac{\partial S_\ell(t,g,c)}{\partial t_j}.$$

with $f_{ix} = \frac{\partial F_i(x_i,r_i)}{\partial x}$, $f_{ir} = \frac{\partial F_i(x_i,r_i)}{\partial r}$, $h_{it} = \frac{\partial H_i(t_i,g_{k(i)},c)}{\partial t}$, and $1[i = j]$ equal to one if the condition in brackets is true (and zero otherwise). Defining the $n \times n$ matrices $\nabla_t S = [\frac{\partial S_i(t,g,c)}{\partial t_j}]$, $\nabla_t E = [1[i = j](f_{ix}h_{it} - f_{ir})]$, and $\nabla P = [p_{ij}]$, we get (in matrix notation)

$$\nabla_t S = \nabla_t E + \nabla P \, \nabla_t S.$$

Assuming $\Pi = (I - \nabla P)^{-1}$ exists (with I the $n \times n$ identity matrix), we have

$$\nabla_t S = \Pi \, \nabla_t E,$$

or spelled out,

$$\frac{\partial S_i(t,g,c)}{\partial t_j} = \sum_{\ell \in N} \pi_{i\ell}(1[\ell = j](f_{\ell x}h_{\ell t} - f_{\ell r})) = \pi_{ij}(f_{jx}h_{jt} - f_{jr}). \tag{4}$$

Second, with respect to group instruction time, we have

$$\frac{\partial S_i(t,g,c)}{\partial g_k} = 1[k(i) = k](f_{ix}h_{ig} - f_{ir}) + \sum_{\ell \in N} p_{i\ell} \frac{\partial S_\ell(t,g,c)}{\partial g_k}.$$

Defining the $n \times m$ matrices $\nabla_g S = [\frac{\partial S_i(t,g,c)}{\partial g_j}]$ and $\nabla_g E = [1[k(i) = j](f_{ix}h_{ig} - f_{ir})]$, we get

$$\nabla_g S = \nabla_g E + \nabla P \, \nabla_g S.$$

We now obtain

$$\nabla_g S = \underbrace{(I - \nabla P)^{-1}}_{\Pi} \nabla_g E,$$

or spelled out,

$$\frac{\partial S_i(t,g,c)}{\partial g_k} = \sum_{j \in N} \pi_{ij} 1[k(j) = k](f_{jx} h_{jg} - f_{jr}) = \sum_{j \in N_k} \pi_{ij} (f_{jx} h_{jg} - f_{jr}). \qquad (5)$$

Third, with respect to classroom instruction time, we have

$$\frac{\partial S_i(t,g,c)}{\partial c} = f_{ix} h_{ic} - f_{ir} + \sum_{j \in N} p_{ij} \frac{\partial S_j(t,g,c)}{\partial c}.$$

Defining the $n \times 1$ vectors $\nabla_c S = [\frac{\partial S_i(t,g,c)}{\partial c}]$ and $\nabla_c E = [f_{ix} h_{ic} - f_{ir}]$, we have

$$\nabla_c S = \nabla_c E + \nabla P \, \nabla_c S,$$

leading to

$$\nabla_c S = \underbrace{(I - \nabla P)^{-1}}_{\Pi} \nabla_c E,$$

or spelled out,

$$\frac{\partial S_i(t,g,c)}{\partial c} = \sum_{j \in N} \pi_{ij} (f_{jx} h_{jc} - f_{jr}). \qquad (6)$$

The first-order conditions of the Lagrangian, with value function

$$V(s) = V(S_1(t,g,c), \ldots, S_n(t,g,c)),$$

are

$$\begin{array}{rcll} \sum_{i \in N} v_i \frac{\partial S_i(t,g,c)}{\partial t_j} - \lambda_b + \lambda_{t,j} & = & 0, & \text{for } j \text{ in } N, \\ \sum_{i \in N} v_i \frac{\partial S_i(t,g,c)}{\partial g_k} - \lambda_b + \lambda_{g,k} & = & 0, & \text{for } k \text{ in } M, \\ \sum_{i \in N} v_i \frac{\partial S_i(t,g,c)}{\partial c} - \lambda_b + \lambda_c & = & 0, & \end{array}$$

with $v_i = \frac{\partial V(s)}{\partial s_i}$ for all i in N.

Using equations (4), (5), and (6), the first-order conditions can be written as

$$\begin{aligned}
(f_{jx}h_{jt} - f_{jr}) \textstyle\sum_{i\in N} v_i\pi_{ij} - \lambda_b + \lambda_{t,j} &= 0, \quad \text{for } j \text{ in } N, \\
\textstyle\sum_{j\in N_k}(f_{jx}h_{jg} - f_{jr}) \sum_{i\in N} v_i\pi_{ij} - \lambda_b + \lambda_{g,k} &= 0, \quad \text{for } k \text{ in } M, \\
\textstyle\sum_{j\in N}(f_{jx}h_{jc} - f_{jr}) \sum_{i\in N} v_i\pi_{ij} - \lambda_b + \lambda_c &= 0.
\end{aligned}$$

B Proof of theorem 2

In case of instruction spillovers, the test scores are defined by

$$s_i = F_i(x_i, r_i),$$

with $x_i = H_i(t_i, g_{k(i)}, c) + P_i(x_1, x_2, \ldots, x_n)$ and $r_i = T - t_i - g_{k(i)} - c$. Let $X_i(t, g, c)$ denote the instruction solution for pupil i. We thus have

$$X_i(t, g, c) = H_i(t_i, g_{k(i)}, c) + P_i(X_1(t, g, c), \ldots, X_n(t, g, c)),$$

for all i in N.

First, for private instruction time, we have

$$\frac{\partial X_i(t, g, c)}{\partial t_j} = 1[i = j]h_{it} + \sum_{\ell\in N} p_{i\ell}\frac{\partial X_\ell(t, g, c)}{\partial t_j},$$

with $h_{it} = \frac{\partial H_i(t_i, g_{k(i)}, c)}{\partial t}$. We can define the $n \times n$ matrices $\nabla_t X = [\frac{\partial X_i(t,g,c)}{\partial t_j}]$, $\nabla_t H = [1[i = j]h_{it}]$, and $\nabla P = [p_{ij}]$, to obtain (in matrix notation)

$$\nabla_t X = \nabla_t H + \nabla P\, \nabla_t X.$$

Assuming $\Pi = (I - \nabla P)^{-1}$ exists (with I the $n \times n$ identity matrix) , we have

$$\nabla_t X = \Pi\, \nabla_t H,$$

or spelled out,

$$\frac{\partial X_i(t, g, c)}{\partial t_j} = \sum_{\ell\in N} \pi_{i\ell}(1[\ell = j]h_{\ell t}) = \pi_{ij}h_{jt}. \tag{7}$$

Second, with respect to group instruction time, we have

$$\frac{\partial X_i(t, g, c)}{\partial g_k} = 1[k(i) = k]h_{ig} + \sum_{\ell\in N} p_{i\ell}\frac{\partial X_\ell(t, g, c)}{\partial g_k},$$

with $h_{ig} = \frac{\partial H_i(t_i, g_{k(i)}, c)}{\partial g}$. Defining the $n \times m$ matrix $\nabla_g X = [\frac{\partial X_i(t,g,c)}{\partial g_j}]$ and $\nabla_g H = [1[k(i) = j] h_{jg}]$, we have

$$\nabla_g X = \nabla_g H + \nabla P \, \nabla_g X.$$

We now obtain

$$\nabla_g X = \underbrace{(I - \nabla P)^{-1}}_{\Pi} \nabla_g H,$$

or spelled out,

$$\frac{\partial X_i(t,g,c)}{\partial g_k} = \sum_{j \in N} \pi_{ij} 1[k(j) = k] h_{jg} = \sum_{j \in N_k} \pi_{ij} h_{jg}. \tag{8}$$

Third, with respect to classroom instruction time, we have

$$\frac{\partial X_i(t,g,c)}{\partial c} = h_{ic} + \sum_{j \in N} p_{ij} \frac{\partial X_j(t,g,c)}{\partial c},$$

with $h_{ic} = \frac{\partial H_i(t_i, g_{k(i)}, c)}{\partial c}$. Defining the $n \times 1$ vectors $\nabla_c X = [\frac{\partial X_i(t,\tau)}{\partial c}]$ and $\nabla_{c-r} H = [h_{ic}]$, we have

$$\nabla_c X = \nabla_{c-r} H + \nabla P \, \nabla_c X,$$

leading to

$$\nabla_c X = \underbrace{(I - \nabla P)^{-1}}_{\Pi} \nabla_c X,$$

or spelled out,

$$\frac{\partial X_i(t,g,c)}{\partial c} = \sum_{j \in N} \pi_{ij} h_{jc}. \tag{9}$$

The first-order conditions of the Lagrangian in equation (1), with value function

$$V(s) = V(F_1(X_1(t,g,c), r_1), F_2(X_2(t,g,c), r_2), \ldots, F_n(X_n(t,g,c), r_n)),$$

are

$$\begin{array}{rcll} \sum_{i \in N} v_i f_{ix} \frac{\partial X_i(t,g,c)}{\partial t_j} - v_j f_{jr} - \lambda_b + \lambda_{t,j} & = & 0, & \text{for } j \text{ in } N, \\ \sum_{i \in N} v_i f_{ix} \frac{\partial X_i(t,g,c)}{\partial g_k} - \sum_{i \in N_k} v_i f_{ir} - \lambda_b + \lambda_{g,k} & = & 0, & \text{for } k \text{ in } M, \\ \sum_{i \in N} v_i f_{ix} \frac{\partial X_i(t,g,c)}{\partial c} - \sum_{i \in N} v_i f_{ir} - \lambda_b + \lambda_c & = & 0, & \end{array}$$

with $v_i = \frac{\partial V(s)}{\partial s_i}$ for all i in N.

Using equations (7), (8), and (9), the first-order conditions can be rewritten as

$$\begin{aligned}
\textstyle\sum_{i\in N} v_i f_{ix}\pi_{ij}h_{jt} - v_j f_{jr} - \lambda_b + \lambda_{t,j} &= 0, \quad \text{for } j \text{ in } N,\\
\textstyle\sum_{i\in N} v_i f_{ix}\sum_{j\in N_k}\pi_{ij}h_{jg} - \sum_{i\in N_k} v_i f_{ir} - \lambda_b + \lambda_{g,k} &= 0, \quad \text{for } k \text{ in } M,\\
\textstyle\sum_{i\in N} v_i f_{ix}\sum_{j\in N}\pi_{ij}h_{jc} - \sum_{i\in N} v_i f_{ir} - \lambda_b + \lambda_c &= 0.
\end{aligned}$$

C Survey questions

In this section, we provide a translation of the survey questions that are relevant for our model.

C.1 Mathematics knowledge and skills assessment

- **Current math level**. If you had to assess the overall mathematics knowledge and skills of your students today, how would you rate each of them? **Answer options**: Very weak, Weak, Rather weak, Rather strong, Strong, Very strong
- **Math level at begin of the year**. How were the mathematics knowledge and skills of your students at the beginning of this school year? **Answer options**: Very weak, Weak, Rather weak, Rather strong, Strong, Very strong
- **Language is a barrier for math**. Indicate for your students whether language skills form a barrier for the subject of mathematics. **Answer options**: Language is not a barrier, Language is a slight barrier, Language is a significant barrier
- **Math score on 100**. If you were to test the overall mathematics knowledge and skills of your students today, what score (a number between 0 and 100) would each of them achieve? You may base this on the report cards from the past school year.
- **Individual instruction learning speed**. Reflecting on the past school year, if you were to explain a mathematics exercise individually to a student, how quickly would each of them master this exercise? **Answer options**: Very slow, Slow, Average, Fast, Very fast
- **Classroom instruction learning speed**. Reflecting on the past school year, if you were to explain a mathematics exercise to the entire class, how quickly would

each of them master this exercise? **Answer options**: Very slow, Slow, Average, Fast, Very fast

- **Self study learning speed**. Reflecting on the past school year, if a student were to work on a mathematics exercise independently, how quickly would each of them master this exercise? **Answer options**: Very slowly, Slowly, Average, Fast, Very fast

- **Quantifying learning speeds**: The following two questions are difficult and hypothetical, but essential for our research. We ask you to answer them as carefully as possible. This question concerns the learning speed of your pupils. Assume you had extra time during the past school year, for example, an additional hour for math lessons every week on Wednesday afternoons. There are three options for how to use this time: **Answer options**: Class instruction time - You address the entire class during the full extra hour (e.g., giving examples or reviewing homework/tests together), Individual instruction time - You focus exclusively on one pupil (or a small group) during the entire extra hour, for example, providing remediation or extra challenges, while the other pupils work independently, Self study - You do not address any pupil directly, allowing all pupils to process the learning material or complete exercises independently.

- How much will a pupil with average learning speed progress in each case? For example, the average learning speed pupil now has a score of 75. After an extra hour of classroom instruction time they will have a score of 75 + X. Then X is what you should fill in at classroom instruction time. **Answer options**: Progress of a pupil with average learning speed in case of classroom instruction time, Progress of a pupil with average learning speed in case of individual instruction time, Progress of a pupil with average learning speed in case of self study.

- Afterwards we show them a matrix with all learning speeds (very slow, slow, average, fast, very fast) and the three modes (class-room instruction, individual instruction, self-study). The average learning speed is already filled in with their previous answer.

C.2 Time use data

- **Class time allocation**. In a typical week, what percentage of class time do you use the following teaching methods: **Answer options**: Individual: addressing a student or a small group of students, e.g., for remediation or extra challenge, while other students work independently, Whole class: addressing the entire class, e.g., to introduce a new concept or work through example exercises, Independent: not addressing any student, with all students working independently.

- **Individual instruction time**. In a typical week, how often do you spend individual time with each of the following students for mathematics? **Answer options**: Never, Almost never, Sometimes, Often, Almost always, Always

- **Minutes per week**. For the previous question about individual time spent on mathematics, how many minutes per week do you have in mind for each of these answer options: **Answer options**: Never, Almost never, Sometimes, Often, Almost always, Always

C.3 Pupil demographics

- Fill in the following basic information for each student:
  - **Grade retention**. Do they have grade retention? **Answer options**: Yes, No
  - **Gender**. What is their gender? **Answer options**: Male, Female, X

- These questions probe the educational disadvantage indicators of your students. Please provide your best intuition.
  - **Dutch at home**. Does this student speak Dutch at home?
  - **Mother has diploma**. Does the mother of this student have a higher secondary education diploma?

  Answer options: Certainly not, Probably not, Probably yes, Certainly yes

C.4 Classroom information

- **Grade**. What grade is your class? (multiple answers possible, e.g., in the case of a mixed-grade class for mathematics) **Answer options**: First grade, Second grade, Third grade, Fourth grade, Fifth grade, Sixth grade

- **Class room set-up**. What classroom setup do you usually use when your students work independently on mathematics? **Answer options**: Everyone at a separate desk, In pairs, with a fixed partner, In pairs, with a rotating partner, In fixed groups, In rotating groups
- **Group size**. You answered that they usually work in groups. How large are these groups? **Answer options**: Size of the smallest group?, Size of the largest group?
- **Main Formation** If you use individual instruction time, do you mainly do this: **Answer options**: One-on-one?, In (small) groups?, Both one-on-one as in (small) groups?

C.5 Teacher background

- Finally, we will ask some questions about you, your teaching career, and your background.
- **Grade experience**. How many years have you been teaching this grade (or these grades)?
- **Primary experience**. How many years have you been teaching in primary education?
- **General experience**. How many years have you been teaching in general?
  - **Gender**. What is your gender? **Answer options**: Male, Female, X
  - **Education**. What is your highest diploma? **Answer options**: Higher secondary education, Professional bachelor, Academic bachelor, Academic master, Other
- Finally, a question about your background when you yourself attended primary school. These questions are again based on educational disadvantage indicators.
  - **Dutch at home**. Did you speak Dutch at home when going to elementary school?

– **Mother has diploma**. Did your mother had a higher secondary education diploma when you went to primary school?

Answer options: Certainly not, Probably not, Probably yes, Certainly yes

D Descriptive statistics for subsamples

This section provides a comparison of the descriptive statistics for the main sample, the subsample used for testing the models (test sample), and the subsample used to compute the marginal social welfare weights (msww sample). The representativeness of these subsamples is assessed by comparing the ratio and distribution of some key variables.

D.1 The test subsample

Table 7: Descriptive statistics for the test subsample

Variable	N	Mean	Std.Dev	Min	Max
Female	1730.0	0.50	0.50	0.0	1.0
Grade retention	1806.0	0.15	0.36	0.0	1.0
Score	1805.0	75.90	18.38	0.0	100.0
Individual time	1806.0	65.56	86.83	0.0	1400.0
Progress individual instr	1806.0	12.02	14.04	0.0	92.0
Progress class instr	1806.0	6.76	9.82	0.0	90.0
Progress self study	1806.0	3.80	10.25	-5.0	90.0
	Weak	Rath. Weak	Average	Rath. Strong	Strong
Math level now	176	269	565	425	371
	(9.7%)	(14.9%)	(31.3%)	(23.5%)	(20.5%)
Math level begin year	217	309	569	377	332
	(12.0%)	(17.1%)	(31.5%)	(20.9%)	(18.4%)
	Cert. no	Prob. no	Prob. yes	Cert. yes	
Mother diploma	113	193	307	897	
	(7.5%)	(12.8%)	(20.3%)	(59.4%)	
Dutch home	337	197	137	973	
	(20.5%)	(12.0%)	(8.3%)	(59.2%)	
	No issue	Small issue	Big issue		
Dutch level	1158	438	210		
	(64.1%)	(24.3%)	(11.6%)		

Table 7 shows the descriptive statistics for the test sample. Overall, this subsample appears to be representative of the main sample, with most variable distributions maintaining similar proportions.

For example, the gender distribution remains balanced, with males accounting for approximately 49% in the main sample and 50% in the test sample. The distribution of grade retention is also similar, with 19.4% of pupils having repeated a grade in the main sample compared to 18.0% in the test sample. However, there are slight deviations in academic performance and study progress measures. Pupils in the test sample show slightly lower average progress across individual instruction (12.02 compared to 13.28) and class instruction (6.76 compared to 7.91). Despite these differences, the overall trends remain consistent.

Notable deviations are observed in the distribution of *math level now* and *math level begin year*. In the main sample, 32% of pupils are classified as "Average" at the current math level, while this percentage drops to 31% in the subsample. Similarly, the proportion of pupils classified as "Strong" decreases from 20% in the main sample to 18% in the subsample.

D.2 The msww subsample

Table 8: Descriptive statistics for the msww subsample

Variable	N	Mean	Std.Dev	Min	Max
Female	1189.0	0.51	0.50	0.0	1.0
Grade retention	1189.0	0.15	0.36	0.0	1.0
Score	1189.0	74.46	18.01	4.0	100.0
Individual time	1189.0	74.31	92.55	1.0	1400.0
Progress individual instr	1189.0	13.15	15.81	0.0	92.0
Progress class instr	1189.0	7.47	11.52	0.0	90.0
Progress self study	1189.0	4.14	11.60	-5.0	90.0
	Weak	Rath. Weak	Average	Rath. Strong	Strong
Math level now	130	199	417	263	180
	(10.9%)	(16.7%)	(35.1%)	(22.1%)	(15.1%)
Math level begin year	166	219	410	225	169
	(14.0%)	(18.4%)	(34.5%)	(18.9%)	(14.2%)
	Cert. no	Prob. no	Prob. yes	Cert. yes	
Mother diploma	94	165	228	702	
	(7.9%)	(13.9%)	(19.2%)	(59.0%)	
Dutch home	203	135	89	762	
	(17.1%)	(11.4%)	(7.5%)	(64.1%)	
	No issue	Small issue	Big issue		
Dutch level	784	261	144		
	(65.9%)	(22.0%)	(12.1%)		

Table 8 presents descriptive statistics for the subsample used to calculate the marginal social welfare weights. This subsample also aligns well with the main sample in most dimensions, but a few variables show noticeable differences.

The gender distribution remains balanced (49% male, 51% female), consistent with the main sample. The distribution of *dutch home* and *mother diploma* categories remains largely stable. However, the percentage of pupils categorized as having "No issue" with Dutch proficiency is slightly higher in the msww sample (66%) compared to the main sample (65%).

There are more pronounced differences in math level distributions. For instance, the percentage of pupils with a "Strong" math level now drops from 20% in the main sample to 15% in the msww sample, suggesting that this subsample may underrepresent higher-achieving pupils. The average individual instruction time is also slightly higher in the msww sample.

E Testing Separability

In our survey, we do not have data about the factors $h_{jt}, h_{jg}, h_{jc}, f_{jg}$ and f_{jx}, but only about the products $h_{jt}f_{jx}$ and $h_{jc}f_{jx}$. In the knowledge spillovers case, we only use the products $h_{jt}f_{jx}$ and $h_{jc}f_{jx}$, leaving us only to estimate $h_{jg}f_{jx}$. We estimate this in the model to be between $h_{jc}f_{jx}$ and $h_{jt}f_{jx}$, without further restrictions. With instruction spillovers, however, we must know the products $h_{jt}f_{ix}$ and $h_{jc}f_{ix}$ to test the model. To proceed, we assume that the ratio $\frac{h_{jc}}{h_{jt}}$ is constant across learning speeds (per class).

To test this separability hypothesis, we compare two models:

1. **Restricted Model:** The ratio is explained solely by teacher-specific effects, that is,
$$\frac{h_{jc}}{h_{jt}} = \alpha + \gamma_j + \varepsilon_{ij},$$
where γ_j represents teacher fixed effects.

2. **Unrestricted Model:** The ratio is explained by teacher fixed effects *and* learning speeds, that is,
$$\frac{h_{jc}}{h_{jt}} = \alpha + \gamma_j + \sum_{k=1}^{4} \delta_k D_{ik} + \varepsilon_{ij},$$
where D_{ik} are dummy variables for the different learning speeds.

The null hypothesis is separability, that is, learning speeds do not affect the ratio:

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0.$$

We compare these nested models using an F-test.[27]

Table 9: ANOVA results with teacher fixed effects

Model	df_resid	SSR	df_diff	SS_diff	F	p-value
Restricted	364.0	33.163193	0.0	—	—	—
Unrestricted	360.0	30.588763	4.0	2.574430	7.574634	0.000007

Table 9 shows the results of the ANOVA test. The unrestricted model (including both teacher fixed effects and learning speed fixed effects) shows a significant improvement in fit over the restricted model (teacher fixed effects only). The F-statistic is 7.575 with a p-value of 0.000007, leading to rejection of the null hypothesis. We can therefore reject separability.

F Time assignment

First, if the teacher indicated that the main classroom setup is primarily one-on-one, we set all group times to zero and allocate one-on-one time to every pupil who receives individual instruction. This applies to five classes.

Second, if the teacher indicated that they mainly work in small groups, we set all group times to a positive number, except for the group of individuals who do not receive any extra time. No one will receive one-on-one time. This applies to 26 classes.

Third, for the last and largest group (68 classes), where the teacher indicated that they work both one-on-one and in groups, we use the following procedure. We first calculate the available disposable time ($t_{\text{disposable}}$) as the product of the number of weekly math hours, the number of minutes per class, and the percentage of individual instruction, divided by 100. Additionally, we calculate T_{group} as the sum of unique t_i values. If $t_{\text{disposable}} \leq T_{\text{group}}$, the individual time is allocated to the pupil with the lowest marginal rate of technical substitution (MRTS, defined as $\frac{f_{ix}h_{jc}-f_{ir}}{f_{ix}h_{jt}-f_{ir}}$) among the pupils with the highest t_i. Specifically, time is given lexicographically to the pupil with the minimum MRTS among those with the maximum t_i. If $t_{\text{disposable}} > T_{\text{group}}$, time is

[27] We use the standard analysis of variance (ANOVA) procedure implemented in Python.

allocated lexicographically by giving it first to the pupils with the highest t_i and second to the pupils with the lowest MRTS. This is intended to represent the least restrictive case.[28] If there is still time left after assigning all time at a given t_i level, the allocation moves to the next highest t_i. Once all pupils at a certain t_i level receive individual time, the group is assigned a strict positive $\tilde{\lambda}_{g,k}$, as no additional time is allocated at the group level.

G Regressions with instruction spillovers

Tables 10 and 11 show the regression results for instruction spillovers with fixed and block peer effects. The magnitude of the coefficients is closer to the regressions without peer effects, indicating that instruction spillovers have a smaller influence on the estimated coefficients compared to knowledge spillovers. As with knowledge spillovers, block peer effects make all estimates insignificant.

[28]Admittedly, this assignment might not be least restrictive, as we could prioritize pupils with the lowest MRTS first.

Table 10: Drivers of the marginal welfare weights (with fixed instruction peer effects)

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5
Math level begin year					
Rather weak	-0.028	-0.028	-0.029	-0.031	**-0.042**
	(-0.062, 0.001)	(-0.057, 0.006)	(-0.061, 0.000)	(-0.066, 0.000)	(-0.079, -0.014)
Average	**-0.047**	**-0.046**	**-0.046**	**-0.055**	**-0.067**
	(-0.093, -0.003)	(-0.091, -0.004)	(-0.090, -0.007)	(-0.112, -0.014)	(-0.113, -0.022)
Rather Strong	**-0.048**	**-0.050**	**-0.048**	**-0.059**	**-0.068**
	(-0.097, -0.003)	(-0.102, -0.002)	(-0.099, -0.003)	(-0.126, -0.012)	(-0.125, -0.022)
Strong	-0.047	**-0.049**	-0.049	**-0.062**	**-0.076**
	(-0.104, 0.019)	(-0.098, -0.003)	(-0.097, 0.016)	(-0.125, -0.004)	(-0.137, -0.011)
Female		0.001	0.002	0.001	0.002
		(-0.018, 0.015)	(-0.015, 0.019)	(-0.016, 0.018)	(-0.018, 0.016)
Dutch home					
Probably no			-0.009	-0.012	-0.011
			(-0.026, 0.010)	(-0.029, 0.006)	(-0.030, 0.007)
Probably yes			0.002	0.005	0.005
			(-0.019, 0.025)	(-0.019, 0.033)	(-0.021, 0.031)
Certainly yes			-0.000	-0.004	-0.002
			(-0.020, 0.021)	(-0.021, 0.017)	(-0.021, 0.015)
Mother diploma					
Probably no				0.025	0.024
				(-0.011, 0.063)	(-0.009, 0.060)
Probably yes				-0.024	-0.024
				(-0.087, 0.029)	(-0.081, 0.025)
Certainly yes				0.021	0.016
				(-0.020, 0.069)	(-0.017, 0.059)
Spline	Yes	Yes	Yes	Yes	No
Teacher Fixed Effects	Yes	Yes	Yes	Yes	Yes

Table 11: Drivers of the marginal welfare weights (with block instruction peer effects)

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5
Math level begin year					
Rather weak	-0.032	-0.031	-0.033	-0.034	-0.043
	(-0.073, 0.003)	(-0.070, 0.005)	(-0.072, 0.001)	(-0.076, 0.002)	(-0.093, 0.001)
Average	-0.050	-0.050	-0.050	-0.057	-0.066
	(-0.121, 0.020)	(-0.119, 0.019)	(-0.119, 0.018)	(-0.129, 0.012)	(-0.161, 0.028)
Rather Strong	-0.056	-0.059	-0.057	-0.066	-0.074
	(-0.147, 0.043)	(-0.149, 0.039)	(-0.149, 0.042)	(-0.161, 0.035)	(-0.191, 0.052)
Strong	-0.061	-0.063	-0.064	-0.074	-0.086
	(-0.171, 0.054)	(-0.166, 0.049)	(-0.171, 0.048)	(-0.183, 0.044)	(-0.204, 0.046)
Female		-0.001	-0.000	-0.000	-0.000
		(-0.021, 0.018)	(-0.020, 0.019)	(-0.020, 0.019)	(-0.021, 0.019)
Dutch home					
Probably no			-0.012	-0.014	-0.013
			(-0.041, 0.018)	(-0.044, 0.013)	(-0.045, 0.016)
Probably yes			0.002	0.005	0.005
			(-0.034, 0.037)	(-0.033, 0.042)	(-0.034, 0.040)
Certainly yes			-0.003	-0.005	-0.004
			(-0.027, 0.021)	(-0.029, 0.019)	(-0.027, 0.020)
Mother diploma					
Probably no				0.020	0.020
				(-0.018, 0.061)	(-0.016, 0.060)
Probably yes				-0.018	-0.018
				(-0.076, 0.032)	(-0.072, 0.030)
Certainly yes				0.015	0.012
				(-0.030, 0.062)	(-0.033, 0.057)
Spline	Yes	Yes	Yes	Yes	No
Teacher Fixed Effects	Yes	Yes	Yes	Yes	Yes