Allocating instruction time: how language instruction can affect multiple skills

Borghans, L.; Diris, R.

Allocating Instruction Time: How Language Instruction Can Affect Multiple Skills

Lex Borghans
Maastricht University

Ron Diris
Katholieke Universiteit Leuven

There exists substantial variation in how schools allocate instruction time to school subjects. The effectiveness of that allocation depends on the immediate effect of instruction in one subject on achievement in the same subject, on how skills further develop over time, and on possible spillover effects on achievement in other subjects. Exploiting a policy intervention in Dutch primary education, we find that effects of language instruction on language skills fade away quickly, while effects of (early) language instruction on several other skills are long-lasting. The results illustrate that spillover effects can arise in the context of skill acquisition.

I. Introduction

What is the value of time spent on education in a certain school subject? The immediate knowledge acquired in that specific subject is an obvious gain, but the long-run value strongly depends on the further development of that skill over time. Moreover, what is learned in one subject can benefit subsequent learning in other subjects. For example, well-developed reading and listening skills can make one more adept at learning and understanding subjects such as geography or history. If one wants to assess the value of instruction in a specific subject, all these effects need to be taken into account.

The aim of this study is to analyze the direct and indirect effects of changes in the instruction time for language on student achievement. We model a skill technology function in which the acquisition of skills depends on initial skills and the allocation of time to all school subjects. Three dimensions of the effect of instruction are crucial: direct effects on the

We would like to thank Thomas Dohmen, Andries de Grip, Olivier Marie, Frederic Vermeulen, and the participants of the European Association of Labour Economists 2011 conference and European Economic Association 2011 conference for their valuable comments.

[*Journal of Human Capital*, 2014, vol. 8, no. 2]
 1932-8575/2014/0802-0003$10.00

specific skill that is being studied, the long-run development of that skill, and effects on other skills. This study uses data from PRIMA, which is a biennial longitudinal survey of Dutch primary schools, executed from the academic year 1994/95 to 2004/5. We estimate both short- and long-run effects of instruction time for language on language achievement, as well as its effects on achievement in mathematics, environmental studies, and nonverbal IQ. To empirically identify these effects, we exploit exogenous variation in instruction time over time, caused by a policy change in the Dutch educational system. We obtain positive short-run effects of language instruction time on language achievement, but these effects fade away quickly over time, while the impact of early language instruction on mathematics is long-lasting. Cross-sectional analyses and value-added estimation underestimate the size of these effects. In this specific setting, the main value of an increase in language instruction lies in spillover effects into other types of skills rather than pure language development. From a broader perspective, this is an illustration that spillover effects can be present in the learning process and that language skills can complement skill acquisition in other subjects.

Figures 1–4 essentially summarize the empirical approach of this study. In the summer of 1998, the Dutch government changed the guidelines for what Dutch primary schools should have taught their students at the end of each grade. The new "targets" contained a stronger focus on language skills at the expense of what were considered "noncore school subjects." Figure 1 shows mean levels of language instruction time by grade and year. We see that, across all grades, the time spent in language classes increases exactly when the policy change becomes effective (in 1998/99). Figure 2 shows that instruction for math remains relatively constant over the same period. The pattern in figure 1 allows us to exploit the dynamics of the effect of language instruction time since different cohorts of students differ in the number of periods in which they are exposed to increased instruction.

Figure 3 shows mean language and IQ scores per grade for multiple panel cohorts. The mean score of the cohort that enters the corresponding grade in 1996 (the latest prepolicy period) is set to zero. The labels below the figure refer to the year in which each cohort entered grade 2. There are two ways to analyze this figure. We can compare scores between cohorts, which shows that postpolicy cohorts perform better in grades 2 and 4. We can also compare within cohorts, which shows that both the 1994 and the 1996 cohorts catch up in achievement exactly at the point where they are first exposed to the policy. Together, this suggests that the policy has strong short-term effects on language achievement. The figure also suggests that there is no long-run effect of receiving more language instruction in earlier grades, as cohorts that are exposed to the policy only in later grades achieve a complete and immediate catch-up upon first treatment. The only main difference between the pattern for language scores

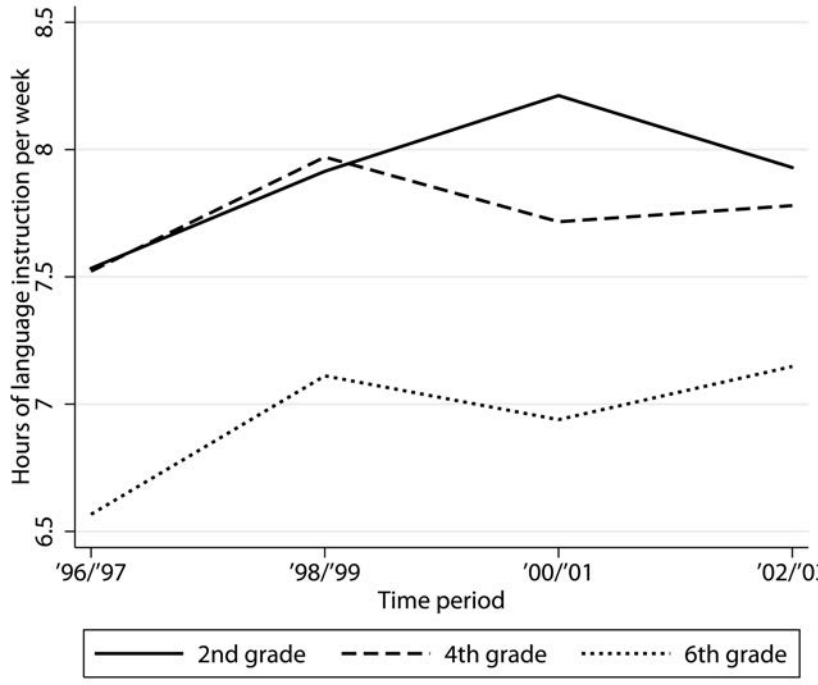

Figure 1.—Mean levels of language instruction time by grade and cohort. The figure shows the mean level of language instruction time in a particular time period and grade for all schools that are present in all four waves of the PRIMA data set. The policy change that we exploit in the empirical analysis first becomes effective in 1998/99.

in the top panel and the pattern for IQ scores in the bottom panel is that the cohort unexposed to the policy in all grades (the 1992 cohort) performs at the same level as other cohorts in language but has lower IQ scores.

Figure 4 shows mean scores for tests in mathematics and environmental studies, which are taken only in grade 6. The figure suggests that there are long-run spillover effects from more early language instruction. The cohort that is exposed to treatment in all grades (2002) performs better on these tests than cohorts exposed only in later grades or not at all. The figure also suggests a negative effect of extra language instruction in grade 6 on math achievement in grade 6; the cohort exposed only in grade 6 (1998) performs worse than the unexposed cohort (1996).

These graphical results are only suggestive and are formalized in the empirical analysis, but they essentially describe the first and second stages of our instrumental variable (IV) estimation. The main identifying assumptions of this IV model are that there are no differences in observable and unobservable characteristics between cohorts and that no other changes that can affect test scores have taken place within the same time frame. We will address each of these issues in the robustness analysis and also

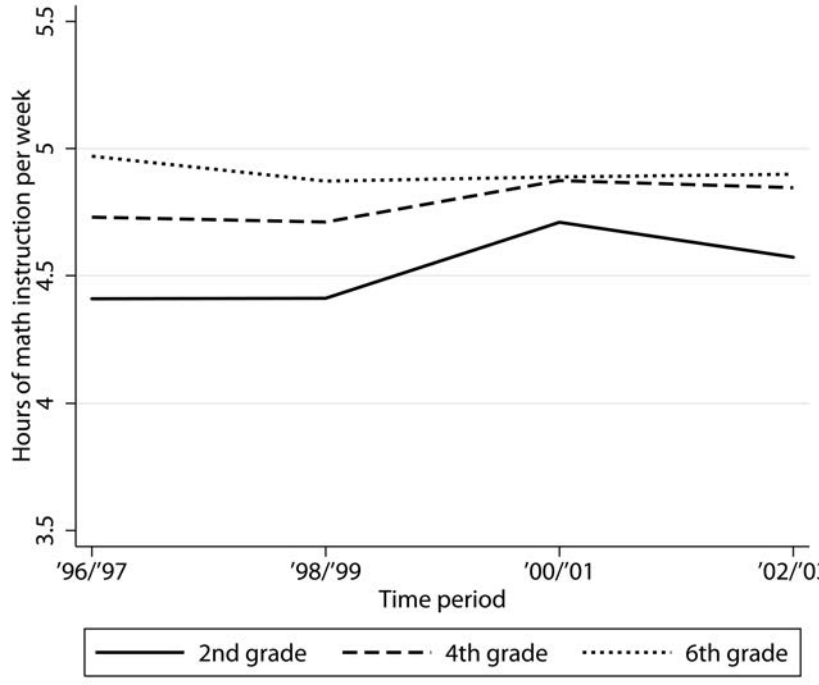

Figure 2.—Mean levels of math instruction time by grade and cohort. The figure shows the mean level of mathematics instruction time in a particular time period and grade for all schools that are present in all four waves of the PRIMA data set. The policy change that we exploit in the empirical analysis first becomes effective in 1998/99.

present results from a value-added model. The conclusions of the value-added analysis are very similar to those of the IV analysis, which suggests that the IV results are not driven by differences in baseline achievement. Additional robustness tests show that it is unlikely that changes at the school or parental level affect our estimates.

Previous research in economics on learning processes in education has largely focused on the added value of (good) teachers. This literature finds that teacher quality matters a great deal but that it is yet unclear to which specific teacher characteristics this can be attributed (see, e.g., Hanushek 1997, 2003; Rivkin, Hanushek, and Kain 2005; Hanushek and Rivkin 2006). Teacher experience or degree explains little of teacher effects. A recent study shows that the conclusions are similar when we look at the added value of principals; they matter for student performance, but the exact pathways are unclear (Dhuey and Smith 2013). Allocating time to school subjects is a key task of schools and teachers and could be part of what distinguishes "good" teachers from "bad" teachers. By assessing how much this allocation matters, we can get insights into part of the "black box" of teacher quality. By assessing both the long-run consequences of extra language instruction and the possible presence of spillovers, we can also get insights into the degree to which the added value of a school can

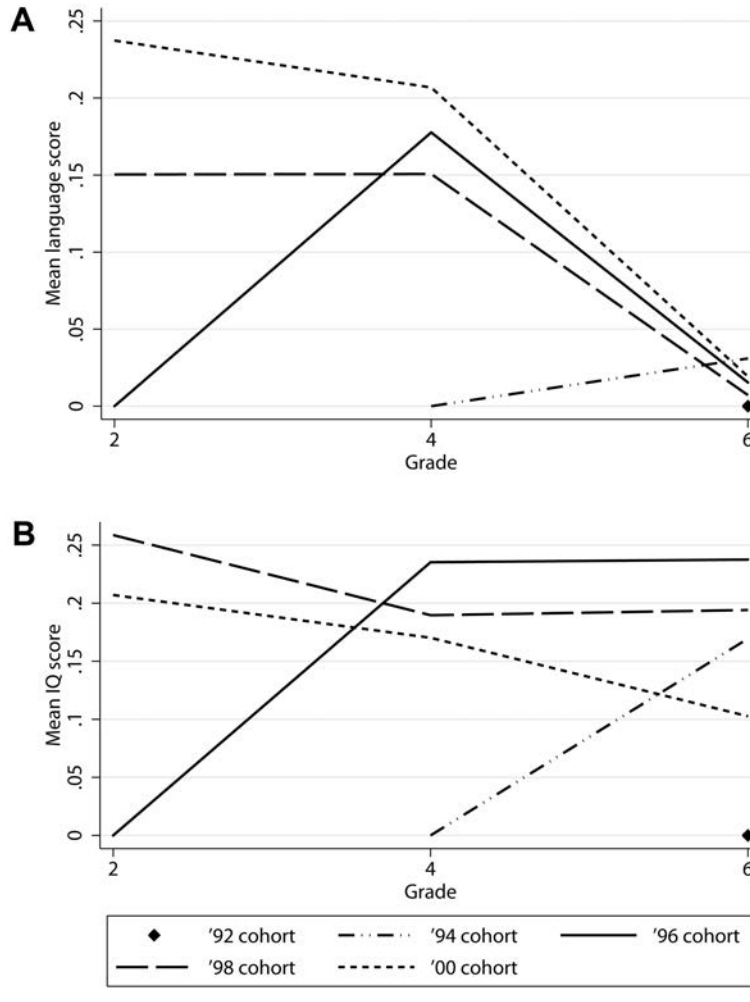

Figure 3.—Mean language and IQ scores by grade and cohort. The figure shows language and IQ scores in all three grades for four separate individual panels of students. The labels refer to the year in which the cohort entered second grade. The scores correspond to mean scores for that specific panel in that specific grade. Scores are standardized with a standard deviation of one, while the mean for the cohort that enters the corresponding grade in 1996 (the latest prepolicy period) is set to zero. There is no observation for the 1994 cohort in grade 2 or for the 1992 cohort in grades 2 and 4 since the data start in 1996.

be seen as the sum of the separate added values of each grade and school subject. Value-added models are often based on this assumption of additivity.

Studies that directly assess the impact of time spent on school subjects are more prominent in educational research (see, e.g., reviews by Sammons [1999] and Scheerens [2000]). These studies generally focus on short-term direct effects instead of taking the perspective of long-term skill acquisition. Moreover, they tend to make use of cross-sectional analyses,

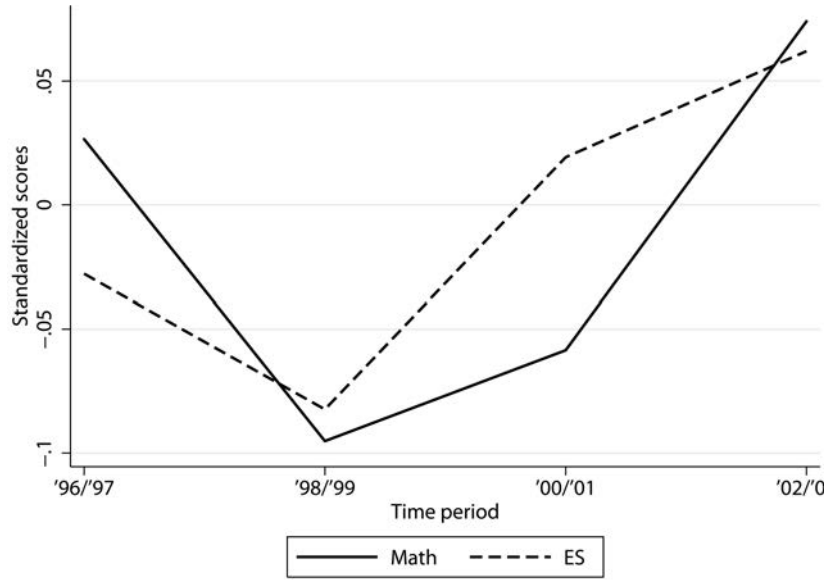

Figure 4.—Mean math and ES scores by cohort. The figure shows mean scores for math and environmental studies in grade 6 for the corresponding cohort. Scores are standardized with mean zero and a standard deviation of one.

which questions the causality of their estimates. Machin and McNally's (2008) quasi-experimental evaluation of the National Literacy Project is one of the few studies on the effects of instruction within the economic literature. Despite this lack of attention, decisions on the content of education are economic decisions. Schools have to make trade-offs with respect to which type of instruction they deem most valuable. Each subject has a specific payoff on one or several skills. Gaining a better understanding of the added values of different school subjects is vital in achieving an efficient allocation of instruction time. From an economic perspective, increases in instruction in a specific subject are justified if the marginal effect on a student's skill set outweighs that of instruction in any other subject. It is crucial that the measure of this marginal effect also looks at skill development in the long run and at achievement in other school subjects. Skills acquired in a certain period are dependent on skills already present before that period. Moreover, if complementarity between subjects is strong, skill acquisition in a certain subject can benefit from proficiency in other skills. This implies that relatively small initial effects on skill acquisition can have strong implications in the long run. As such, this topic fits well within the technology of skill formation as developed by Cunha and Heckman (2007), which treats the development of skills as a dynamic process that depends on initial skills and investments in each period.

The paper is organized as follows. Section II provides the theoretical framework for assessing the impact of changes in instruction time. Section III discusses the PRIMA data sample that we use for our analysis and

also provides background information on the policy change we exploit in the empirical analysis. Section IV presents the methodology for the IV estimation. Section V discusses our main results. Robustness analysis is discussed in Section VI. Section VII presents conclusions.

II. Theoretical Framework

Previous research on the effect of class instruction on student achievement has focused on both the effects of (effective) time spent in class and the effects of specific methods of instruction.[1] The former group of studies provide inconsistent conclusions. Evaluations of the length of the school day or school year find no statistically significant relationship with academic achievement (Harnischfeger and Wiley 1976; Stallings, Needels, and Stayrook 1979). Research on the effects of instruction time in specific subjects on academic achievement generally identifies positive and statistically significant effects, although there are inconsistencies across subjects or grades (Jacobson 1980; Daniels and Haller 1981). Bell and Davidson (1976) even find negative correlations for some subsamples. Cooley and Leinhardt (1980) and Wang (1998) show that opportunity to learn and content exposure are the biggest predictors of test scores from all "classroom information."

This literature almost exclusively focuses on one-to-one correlations between instruction time and test scores in a specific grade. We take a broader approach by also examining long-run effects of instruction time and assessing achievement in multiple subjects. Economic theory tells us that an optimal allocation of instruction time depends on the initial productive value of (instruction in) each subject on the skill being taught, how this value changes over time, and the extent to which the acquired skills in one subject have spillover effects on skills in other subjects.

A. *Long-Run Effects*

We approach the concept of instruction time as an investment in skills, analogous to Cunha and Heckman's (2007) "Technology of Skill Formation." This function specifies the development of skills over the life cycle as a function of the skill set already present (θ), parental characteristics (h), and investments (I):

$$\theta_{it+1} = f_t(\theta_{it}, h_i, I_{it}).$$

Instruction time in a specific school subject is an example of a potentially productive investment in skills. The long-run value of investments in skills depends on the initial investment productivity but also on how skills further develop over time. The latter is determined by the degree of self-productivity of skills and the degree of complementarity between skills

[1] For example, individual or group, written or oral, involved or engaged (see Powell 1978; Stallings 1980).

and investments (Cunha, Heckman, and Schennach 2010). The value of additional instruction time can increase over time when it enhances the learning of new topics and thereby also makes future investments more productive. When skills that are acquired early have little to no value in later grades and complementarity is low, the impact of more instruction can fade away over time. Complementarity could be so low that it does not matter when investments are made, as long as they take place at some point in time. In that case, investments are perfect substitutes, and the learning process is purely additive. Since we conduct our analysis at three different age levels, we can get an indication of the degree of self-productivity and complementarity.

We assume that instruction time (IT) is the only productive investment and ignore parental characteristics here.[2] The skill set θ in the current period t depends on skills already developed in previous periods by a coefficient α_t and on investments in instruction time by a coefficient β_t. We examine three different time periods, grades 2, 4, and 6 (these are the grades we have data on):

$$\theta_2 = \alpha_2\theta_0 + \beta_2 IT_2,$$
$$\theta_4 = \alpha_4\theta_2 + \beta_4 IT_4,$$
$$\theta_6 = \alpha_6\theta_4 + \beta_6 IT_6.$$

We define $\gamma_{t1,t2}$ as the effect of IT in period $t1$ on the skill set in period $t2$. The effect of second-grade IT on skills in grade 2 becomes

$$\gamma_{2,2} = \beta_2.$$

For grade 4, we can estimate the effect of IT in that period but also of IT in the previous period (i.e., grade 2):

$$\gamma_{4,4} = \beta_4,$$
$$\gamma_{2,4} = \beta_2\alpha_4.$$

For grade 6, we can retrieve the effect of IT on skill formation in the current period and in two previous periods:

$$\gamma_{6,6} = \beta_6,$$
$$\gamma_{4,6} = \beta_4\alpha_6,$$
$$\gamma_{2,6} = \beta_2\alpha_4\alpha_6.$$

In this model, α_n captures both the degree of self-productivity and the degree of complementarity. We can specify these effects separately by introducing an interaction term between IT and incumbent skills. In this model, δ measures self-productivity and τ measures complementarity between skills and investment:

[2] This implies that a proper empirical analysis should take into account the possible confounding effects of these factors.

$$\theta_4 = \delta_4\theta_2 + \beta_4 IT_4 + \tau_4\theta_2 IT_4,$$
$$\theta_4 = (\delta_4 + \tau_4 IT_4)\theta_2 + \beta_4 IT_4,$$
$$\theta_4 = \alpha_4\theta_2 + \beta_4 IT_1.$$

Although we can depict this theoretically, it is not possible to separately identify both these effects in our empirical analysis given the data at hand.[3] Whether the estimated impact of IT in a given grade increases or decreases over time depends on whether α_n is larger or smaller than one, respectively.

B. Spillover Effects

The empirical analysis of this study does not assess an extracurricular program, but looks at changes in language instruction within the typical school week. This means that increases in IT for one school subject take time away from other subjects. Effects will be lower than when we would assess the impact of language IT in isolation, if the increase in IT comes at the expense of another subject that is productive for that skill. This does not limit the relevance of our findings. Any increase in IT for a specific subject comes at the expense of instruction in another subject if we hold the length of the school week fixed.[4] However, we have to be aware that the implications of such an increase depend on where the extra time is taken from in that specific situation. Moreover, one needs to know the productive value of instruction in other subjects before positive findings are used as an argument to increase IT for that particular subject.

The fact that increases in IT in one subject come at the expense of IT in other subjects is especially relevant in the presence of spillover effects between different types of skills. When less instruction is received in a subject that creates spillover effects for the skill we measure, the estimates will decrease. Moreover, when we directly assess spillover effects from, for example, language instruction to math achievement, the change in math IT is of especial importance. One of the goals of this paper is to assess whether language instruction has value for skills other than language. Language has an especially high potential for spillover effects since its skills can be applied in a wide range of subjects. Possible spillover effects for language IT might not generalize toward other school subjects, although we should always account for the possibility of their existence in this type of analysis.

Hence, the estimates of the impact of language instruction depend not only on the value of language but also on where any extra time is retrieved and on the productive value of that particular activity. The "underestimation" of the estimate is equal to the product of the two, summed for all

[3] Since all these variables are endogenous, this would require a multitude of instruments and hence very rich data.

[4] One could argue that this also applies to increases in the length of the school day, as it can come at the expense of lower concentration levels or fewer extracurricular activities.

school subjects. We define the degree to which instruction in any other subject decreases as σ_n. This decrease is measured relative to the increase in instruction for language, so that all σ's add up to one. Changes in language instruction (ΔIT^L) affect the levels of instruction in other subjects:

$$\Delta IT^L = -\Delta IT^M - \Delta IT^Q - \cdots - \Delta IT^Z.$$

The superscript M refers to mathematics and the superscripts Q–Z refer to all other school subjects. We specify how much time each subject loses by applying σ's as weights:

$$\Delta IT^M + \Delta IT^Q + \cdots + \Delta IT^Z = -\sigma_1 \Delta IT^L - \sigma_2 \Delta IT^L - \cdots - \sigma_n \Delta IT^L,$$

with $\sigma_1 + \sigma_2 + \cdots + \sigma_n = 1$.

We define the productive value of instruction in every subject n for skills in every subject Z as ω_n^Z. The value added in language skills can then be represented as[5]

$$\begin{aligned}
\Delta\theta^L &= \beta^L \Delta IT^L + \omega_1^L \Delta IT^M + \omega_2^L \Delta IT^Q + \cdots + \omega_n^L \Delta IT^Z,\\
\Delta\theta^L &= \beta^L \Delta IT^L - \sigma_1\omega_1^L \Delta IT^L - \sigma_2\omega_2^L \Delta IT^L - \cdots - \sigma_n\omega_n^L \Delta IT^L,\\
\Delta\theta^L &= (\beta^L - \sigma_1\omega_1^L - \sigma_2\omega_2^L - \cdots - \sigma_n\omega_n^L)\Delta IT^L.
\end{aligned}$$

Similarly, the effect of changes in language instruction on mathematical skills is equal to

$$\Delta\theta^M = (\beta^M - \sigma_1\omega_1^M - \sigma_2\omega_2^M - \cdots - \sigma_n\omega_n^M)\Delta IT^L.$$

To simplify, we specify the total negative component of increases in language instruction, resulting from changes in other IT variables, as λ. It is a weighted average of the added values of all other school subjects, with σ_n applied as weights:

$$\lambda^Z = \sigma_1\omega_1^Z + \sigma_2\omega_2^Z + \cdots + \sigma_n\omega_n^Z.$$

Hence, the effect of IT for language in each grade on language skills in grade 6 is equal to

$$\begin{aligned}
\gamma_{6,6}^L &= \beta_6^L - \lambda_6^L,\\
\gamma_{4,6}^L &= (\beta_4^L - \lambda_4^L)\alpha_6,\\
\gamma_{2,6}^L &= (\beta_2^L - \lambda_2^L)\alpha_4\alpha_6.
\end{aligned} \tag{1}$$

Changes in IT can also induce behavioral changes by students, parents, or teachers, who might adjust their effort levels in response to changes in IT. For example, reductions in IT for language can lead to compensating investments from parents. If such changes are present, the estimated effect

[5] Every β, σ, and λ can also differ by time period, but this is not denoted for ease of notation.

becomes a net result of changes in IT and direct responses from other agents to the treatment.[6]

To achieve an optimal allocation of IT, a policy maker needs to know the (marginal) value of all school subjects in every grade, which allows for calculation of every $\beta_t^Z - \lambda_t^Z$, and he or she also needs to understand the behavioral changes that changes in IT can induce. The focus of our empirical analysis is on IT for language. The model implies that the value of the specific instruction that one assesses has to outweigh the weighted average of the values of instruction in those subjects that suffer decreases in instruction if one is to identify a positive effect. Naturally, λ_t^L is not constant across situations since it depends on the specific allocation of IT across subjects and on the level of other inputs. In fact, β_t^L is influenced by the current level of inputs as well. Although this limits the external validity of our empirical findings, the empirical results from this study can still provide valuable insights into how the relation between IT for language and skill acquisition can work in a specific setting.

III. Policy and Data

A. *The Policy Change*

Dutch primary schools have a lot of freedom in designing their typical weekly curriculum. There are no strict demands with respect to the time spent on certain subjects, but there are demands on what pupils should be able to master at the end of the year. These are specified in the so-called core targets, which enlist the skills students should comprehend for every grade and every subject. Schools are evaluated on these targets, and therefore, they have a guiding role toward the setup of the weekly curriculum, albeit indirect. In the beginning of the 1990s, multiple new topics were brought into the list of demands, largely related to the building of citizenship.[7] In the same period, an independent commission of educational experts assigned by the Dutch government executed an elaborate investigation of the Dutch primary school system. The formation of this Commissie Evaluatie Basisonderwijs (CEB) was largely a response to the fact that large and thorough evaluations of school systems that are so common in the United States were absent in the Netherlands up to that point. The commission gathered test results of students over the period 1988–94, based on tests virtually identical to those used in the empirical analysis of this paper. The final report from the CEB argued that the expansion of targets had taken too much attention away from core activities, leading to a decrease in especially language achievement (Commissie Evaluatie Basisonderwijs 1994). In 1995, a follow-up investigation evaluated the core targets in light

[6] One could simply expand model (1) by a parameter reflecting such responses. We will not do so explicitly here but will address other changes in inputs in the empirical application.

[7] Children learn about social norms, the core values of a democracy, environmental awareness, etc.

of the poor evaluation results from the report. It reached similar conclusions, after which both these pieces of advice were translated into an effective change in policy, namely, the formation of new core targets in 1998. These new targets were more demanding toward language achievement, while the chapter relating to citizenship was reduced and made voluntary.

Schools were not given strict orders to increase the quantity of instruction time for language, but these new guidelines have removed schools and teachers from obligations with respect to "secondary activities." This has freed up instruction time that largely should have been directed toward language, given the emphasis in the new targets on especially this subject. Not every single school will pick up on this policy change to the same degree, but one would expect that this has raised the time spent on language for the average school. If the increase is sufficiently strong, the policy change can be exploited to empirically assess the effect of instruction time for language on school achievement.

B. PRIMA Data

The Dutch PRIMA study assesses primary school students in the Netherlands in a period that envelops this policy change, namely, 1994/95 to 2004/5. PRIMA collects information on students in kindergarten and in second grade, fourth grade, and sixth grade of primary education. Tests are conducted for language, mathematics, and IQ. The IQ test is a nonverbal test, consisting of two parts: combining (putting two figures together to form a prespecified shape) and excluding (selecting which figure is different from the other three).

The language and IQ tests are identical in every wave. The math test contains different questions and also has a different number of questions per wave, which makes scores incomparable. Since we need comparable test scores over time to properly exploit the policy change, we have to rely on an alternative measure of math achievement. PRIMA also includes results for the Dutch Cito test. This high-stakes test of cognitive ability is executed at the end of primary school in the Netherlands (grade 6) and partly determines into which track students are selected in secondary school. The data report an overall Cito score as well as the number of correctly answered questions for the subdomains of math and environmental studies (ES). ES combines biology, geography, and history. Although the questions of the high-stakes test are not identical for every year, they are selected to have the same difficulty level on average, and the total number of questions is identical in every wave. To sum up, we have comparable test scores for language and IQ in grades 2, 4, and 6 and for mathematics and environmental studies in grade 6.[8]

[8] The results for the Cito test for language are not comparable since the number of questions in the test was changed between waves 4 and 5. We include both types of language tests when we estimate the value-added model in Sec. VI.B. PRIMA also reports test scores for reading, but the reading test was not conducted in the wave before the policy change.

PRIMA also administers background questionnaires for pupils, parents, teachers, and principals. Teachers in grades 2, 4, and 6 are asked how much time they spend per week on each school subject. The exact question reads, "How much school time did the pupils, on average, receive per week in the following subjects?" This applies only to language, reading, and math in the first two waves but is expanded with other subjects in later waves. Teachers can answer the question to the exact minute, but we rescale these values to hours. Thus, the values of IT in the remainder of this paper refer to hours of instruction time, but these are not (necessarily) integer numbers. We sum up the time devoted to both reading and language instruction for our measure of IT since both these subjects aim to improve the understanding of language.[9]

The first PRIMA wave (1994/95) differs from the others because tests are conducted at the beginning of the academic year, while this is around March for later waves, and questions concerning instruction time are aimed at the previous year. This means that scores are lower (pupils have had less schooling) and IT values are from another grade (first, third, and fifth). The first PRIMA study is therefore discarded. Although the 1996/97 wave is the earliest study included in the analysis, we refer to it as the second wave in the remainder of this paper. As an abbreviation, we label this wave as T2. The other waves are similarly labeled as T3, T4, T5, and T6.

PRIMA also tests students in kindergarten on basic arithmetic and language skills, but both tests change after wave 2. On the other hand, kindergarten pupils are tested during the same time period in all waves, including wave 1.[10] Hence, waves 1 and 2 are fully comparable, which allows us to control for baseline performance in a restricted analysis.

IV. Methodology

A. Instrumental Variable Model

The allocation of instruction time to different school subjects is decided by schools and teachers and thus is potentially endogenous. IT could correlate strongly with the ability level of the average student in the school and with other school, teacher, and parental inputs as well. This implies that ordinary least squares (OLS) estimates of language IT can be strongly biased. There is large variation in how previous research on the effects of IT deals with selection bias. Estimates are based on simple correlations (Bell and Davidson 1976; Daniels and Haller 1981), OLS including baseline test scores (Cooley and Leinhardt 1980; Kiesling 1984), or hierarchi-

[9] The coefficients in the empirical analysis are strongly similar when we exclude time spent for reading, which is to be expected since the increase in IT was largely the result of more language instruction rather than reading instruction.

[10] In fact, the reason why the waves are not comparable in other grades is that there were practical problems with the kindergarten test in wave 1, after which it was retaken half a year later. For comparability, all other grades were also moved to March from wave 2 on.

cal linear models (Wang 1998). A more recent study by Machin and McNally (2008) uses a quasi-experimental setup to evaluate the use of the National Literacy Project (NLP), which involved more focused literacy instruction and effective classroom management. They compare schools that used NLP with schools that did not in geographically adjacent areas.

We assess the relationship between IT in language and measures of ability and socioeconomic status, that is, education level of the mother and the father and the average overall Cito score of the school over all waves. Table 1 shows a negative and statistically significant correlation between instruction time for language and these indicators. This relation is strongest for grade 2 and weakest for grade 4.[11] Assuming that the association with unobservable variables that also affect test scores is similar, the effect of language IT on achievement will have a negative bias when estimated cross-sectionally. This could explain why many studies find very low or statistically insignificant effects of IT on achievement.

Given that IT is strongly related to other determinants of test scores, of which some are likely unobservable, instrumental variable (IV) estimation becomes a viable alternative. A proper instrument correlates with IT, but not with other determinants of achievement. The educational policy change from 1998 is a suitable candidate. Figure 1 has shown that it led to an increase in IT across all grades. Provided that the increase is strong enough, we can exploit this policy change in an IV model. More specifically, we can employ as an instrument a dummy (D) that takes the value zero for the 1996/97 study and the value one in the cohorts thereafter. This presents us with the following model:

$$IT_i = \alpha_0 + \alpha_1 D_i + \epsilon_i,$$
$$\text{Score}_i = \beta_0 + \beta_1 IT_i + \eta_i.$$

We estimate effects for each specific grade separately. Thus, we are not following the same individuals over time (as these naturally advance in grade) but are comparing the same grades in the same schools in different periods. This sample is essentially a school panel but still contains observations on an individual level. Our dependent variable Score_i can refer to scores in language, IQ, mathematics, and environmental studies. The β_1 in the model above corresponds to $\beta_t^L - \lambda_t^L$ from model (1). It incorporates both the added value of having more IT for language and the negative component of having less IT for other subjects.[12]

[11] The Cito scores can be affected by language IT as well. Assuming a positive true impact of IT, this will only lead to an underestimation of the already large estimates in table 1.

[12] This is presumably very small for the test domains we are looking at, as we explain in the next section.

TABLE 1
Association between Ability Indicators and Instruction Time for Language

Grade Sample	Average Cito Score	Educational Level of Father	Educational Level of Mother
2nd grade	−.029**	−.574***	−.604***
	(.012)	(.112)	(.108)
4th grade	−.022**	−.426***	−.437***
	(.011)	(.094)	(.096)
6th grade	−.038***	−.545***	−.497***
	(.010)	(.086)	(.088)

Note.—Huber-White's robust standard errors are reported in parentheses, corrected for clustering at the school level. In the first column, number of hours of language instruction is regressed on the school average score for the Dutch Cito test, averaged over all waves of the PRIMA data. In the last two columns, number of hours of language instruction is regressed on the school average education level of either the mother or father, averaged separately per wave. Educational level is categorized into four categories, from low to high.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

The fact that language IT increases simultaneously for all grades (see fig. 1) implies that there are differences in the intensity of treatment across cohorts. Some students are treated in no grade and some in all grades; one group of students is treated in grade 6 but not in grades 4 and 2; and another group is treated in both grades 6 and 4 but not in grade 2. This variation allows us to simultaneously estimate short- and long-run effects of language IT for achievement in higher grades. We therefore include lagged IT variables in the regressions for fourth- and sixth-grade achievement. These lags refer to the received levels of IT for that individual 2 years and two grades earlier in the case of the first lag and 4 years and four grades earlier in the case of the second lag. We employ lagged policy dummies as instruments for the lagged IT variables.[13] For the first lag, the instrument has a value of zero for the first two time periods and a value of one for the last two periods. For the second lag, the instrument has a value of zero for the first three periods and a value of one for the last period (see table 2):

$$IT_i^6 = \alpha_0 + \alpha_1 D6_i + \alpha_2 D4_i + \alpha_3 D2_i + u_i,$$
$$IT_i^4 = \gamma_0 + \gamma_1 D6_i + \gamma_2 D4_i + \gamma_3 D2_i + v_i,$$
$$IT_i^2 = \delta_0 + \delta_1 D6_i + \delta_2 D4_i + \delta_3 D2_i + w_i,$$

$$\text{Score}_i = \beta_0 + \beta_1 \hat{IT}_i^6 + \beta_2 \hat{IT}_i^4 + \beta_3 \hat{IT}_i^2 + \epsilon_i. \quad (2)$$

The above pertains to estimation in grade 6; $D6$, $D4$, and $D2$ are the (dummy) instruments for IT in language in grades 6, 4, and 2, respectively.

[13] We do not have lagged values for the earliest cohorts and assume these to be equal to the value of IT for that particular school in wave 2. This refers to the first lag of IT for fourth and sixth graders in T2 and the second lag of IT for sixth graders in T2 and T3.

TABLE 2
Values of Dummy Instruments

	1996/97	1998/99	2000/2001	2002/3
*D*6	0	1	1	1
*D*4	0	0	1	1
*D*2	0	0	0	1

Note.—The table shows the values of the dummy instruments in every period for all three grades. The above refers to estimation for sixth-grade achievement. The dummy *D*6 is the instrument for language instruction in grade 6, *D*4 for that in grade 4 (2 years earlier), and *D*2 for that in grade 2 (4 years earlier). The first two rows are the same for grade 4 achievement but apply to instruction in grades 4 and 2 there. For grade 2 achievement, only the first row applies.

For statistical reasons, we include all instruments in each of the first-stage regressions, but each variable is essentially instrumented by the instrument on the diagonal.[14] For estimation in grade 4, the second lag is absent and only two instruments are included. For grade 2, no lags are present and only one instrument is included. The effect of lagged IT depends on the initial effect of language instruction in combination with the degree of complementarity and self-productivity within and between skills (α_t in model [1]).

B. First-Stage Results

The validity of this model depends, first of all, on the impact of the policy change on the level of IT for language. Figure 1 has shown that IT for language increases directly after the policy change in 1998 in all grades. The magnitude of the increase is around half an hour of extra language instruction, which corresponds to almost half a standard deviation. For the data in figure 1, we include only schools that recur in every wave to increase comparability.[15] Both the figure and the empirical analysis are based on a sample that runs up to wave 5 since this increases the total sample size. This involves dropping wave 6, but we also retrieve schools that are in all waves but 6. The trade-off in terms of total observations is favorable toward including fewer periods. Estimating long-run effects requires at least four waves to ensure that we have variation in the value of all three instruments, but not necessarily a fifth period.[16]

The jump in IT from the 1996/97 cohort to the 1998/99 cohort should ensure a strong correlation between the policy dummies and IT. This is confirmed by the first-stage results from table 3. We report *F*-statistics for the Kleibergen-Paap weak identification test (Kleibergen and Paap 2006),

[14] Hence, $D6_i$ for IT_i^6, $D4_i$ for IT_i^4, and $D2_i$ for IT_i^2. The coefficients for the other dummies are much lower than for the dummy on the diagonal in all cases; see table 3.

[15] The increase in language IT in grades 4 and 6 is less apparent for the sample as a whole. The reason is that most schools that enter the sample in wave 3 have low levels of language IT.

[16] The estimates from the empirical analysis are very similar when we include wave 6, although they are slightly less precisely estimated.

TABLE 3
First Stage: Effect of Policy Dummy on Language IT for Different Samples

	Language Sample	IQ Sample	Math Sample	ES Sample
2nd grade:				
Dummy post-1998	.482***	.482***	. . .	. . .
	(.058)	(.058)		
KP *F*-statistic	68.99	69.11	. . .	. . .
n	3,636	3,634		
4th grade:				
Dummy post-1998	.499***	.500***	. . .	. . .
	(.067)	(.067)		
First lag dummy	.451***	.447***	. . .	. . .
	(.061)	(.061)		
KP *F*-statistic	20.00	20.11	. . .	. . .
n	3,366	3,364		
6th grade:				
Dummy post-1998	.558***	.557***	.732***	.613***
	(.065)	(.065)	(.099)	(.102)
First lag dummy	.668***	.669***	.542***	.325***
	(.069)	(.069)	(.109)	(.097)
Second lag dummy	.684***	.684***	.656***	.765***
	(.085)	(.085)	(.109)	(.127)
KP *F*-statistic	23.44	23.39	15.26	14.90
n	3,101	3,100	1,675	1,328

Note.—Huber-White's robust standard errors are reported in parentheses, corrected for clustering at the school level. The table shows results of the first stage of model (2), estimating the effect of the postpolicy dummy on instruction time for language. We report separate results for each outcome variable that is used in the second stage of the two-stage least squares model, as the number of available observations differs across test scores. The first and second lag dummies measure whether the cohort experienced the policy change, respectively, 2 and 4 years earlier. KP *F*-statistic reports the first-stage *F*-statistic of the Kleibergen-Paap test of weak identification, conducted on the excluded instruments.
* $p < .10$.
** $p < .05$.
*** $p < .01$.

conducted on the excluded instruments. This is the appropriate test statistic for models with multiple endogenous regressors and robust standard errors. We report first stages separately for each outcome variable that is used in the second stage of the IV analysis, as the high-stakes tests for math and ES contain missing observations and the estimates are therefore based on a smaller sample in that case. The *F*-statistics are all above the informal critical value of 10, as proposed by Staiger and Stock (1997).[17]

Instrument validity also depends on the absence of a correlation between the instrument and the error term. This assumption can be threatened when pre- and postpolicy cohorts differ at the baseline or when other

[17] For grades 2 and 4, they are also above the critical values as reported by Stock and Yogo (2005), which are 16.38 for grade 2 estimation and 7.03 for grade 4 estimation. Stock and Yogo do not report critical values for the case of three endogenous regressors and three excluded instruments, which applies to our grade 6 estimation, but the *F*-values are also above the more informal critical value of 10 in that case.

changes have been made within the same time window. We ensure greater comparability by including only schools that appear in every wave. But cohorts within each school can still differ in observable and unobservable characteristics from period to period. We mainly address this in Sections VI.A and VI.B, where we control for observable characteristics and kindergarten test scores and also add estimation of a value-added model. We believe that the nature of the policy change already reduces concerns about changes over time. The policy change was by no means drastic, but by shifting the emphasis in the guidelines, it still increased the average time spent on language without prescribing any qualitative changes. Additionally, because the extra emphasis on language in the educational guidelines came at the expense of IT for improving citizenship, time spent on school subjects that are likely to affect achievement in the subjects we measure is expected to remain constant (we assess this in Sec. IV.C). Finally, because of the time lag between the poor evaluation results and the effective change in policy, strong behavioral changes on the school and parental levels as a response to poor evaluation would have taken place before. We will assess in Sections VI.E and VI.F whether changes at the school level, including the quality and nature of instruction, or at the parental level can confound our estimates.

C. Instruction Time in Other Subjects

As specified in Section II, it is crucial that we know how IT has developed for subjects other than language. The development of IT for math is especially important given that we employ math achievement as one of our outcome variables. Changes in math IT are incorporated in λ_t^L and can strongly affect our estimates. Figure 2 shows that volatility in math IT is very low, especially between the periods before and after the policy change (T2 and T3). The only nonnegligible difference occurs between the third and fourth waves in grade 2, where IT for math increases by 19 minutes per week. We estimate effects for math only at the sixth-grade level, while those who are in grade 2 in waves 4 and 5 never appear in grade 6 in the data. This increase could put an upward bias on the effects of language IT in grade 2 on IQ scores in grades 2 and 4, however. All other variation is lower than 10 minutes of instruction per week. Thus, σ_1 lies close to zero.

Where, then, did this extra language instruction come from? Increases in language tend to come at the expense of, foremost, "all other time."[18] This is defined as the difference between the sum of all reported subjects and the length of the schooling week and consists of activities that could

[18] This is largely based on information from later waves, as they include data on subjects other than math and language. The results from later waves might not automatically generalize to the period after the policy change, but the decision process is believed to be the same; schools feel that they need to spend more time on language, and the new targets have a guiding role in where this time will be taken from.

not be grouped into any of the other eight categories (language, math, environmental studies, physical education, expressive subjects, music, English, and computer science) on which the quantity of IT was reported.[19] The policy change led to targets with a stronger focus on language and reduced demands on the "building of citizenship." It is likely that the bulk of this other time went into those activities, also given the rise in importance of this topic in the period before the policy change. This time is presumably less productive in the subjects for which achievement is tested, and therefore, we would assume λ_t^L to be low.

V. Results

A. *Language Achievement*

Column 1 of table 4 shows the OLS estimates of the effect of language instruction time on language achievement for all three grades. Test scores are standardized with a standard deviation of one. Estimates have a positive sign for all grades, but only the effect for grade 4 is statistically significant. This is likely related to the weaker selection bias in that grade (see table 1). These effects are very modest in magnitude given that they show the effect of an increase in IT of 1 hour each week.

Column 2 of table 4 shows the estimates for the IV model. All short-run effects are substantially larger than in the OLS estimation. This suggests that unobserved heterogeneity, and possibly measurement error in the explanatory variable as well, downwardly bias the OLS estimates. The Durbin-Wu-Hausman test (Hausman 1978) confirms that there is a statistically significant difference between the OLS and IV estimates. The direct effect of language instruction on language achievement is very strong in grade 2. For grade 4, we observe positive and statistically significant effects for current IT (meaning IT received in that particular grade) but statistically insignificant results for lagged IT. A similar picture occurs in grade 6. Current IT is positive and statistically significant, while both lags are statistically insignificant with a negative sign. Hence, increasing IT for language has only a temporary "boosting" effect but no long-lasting positive payoffs for language achievement. These results largely confirm the conclusions from figure 3, which showed that cohorts exposed to the policy change in lower grades had higher test scores in the short run but were caught up by earlier cohorts in later grades. The only main difference with the figure is that in table 4 we also identify a positive short-term effect of IT in grade 6. Overall, the magnitudes of the short-term effects appear to be rather high, but, on an average of around 7.5 hours, 1 hour of extra language instruction every week represents an (economically) significant increase as well. As the

[19] The negative correlation between "other time" and language IT is strongly apparent in all three grades. Statistically significant correlations are also identified between language instruction and instruction in religion (all grades) or English (grade 6), but the strength of these correlations is much weaker.

TABLE 4
Effect of Instruction Time (IT) for Language on Test Scores in Language

Explanatory Variable	OLS (1)	IV (2)	IV (3)	IV (4)	DWH (5)
2nd grade:					
IT	.019	.392***	.351***	.337***	.000
	(.021)	(.093)	(.102)	(.101)	
4th grade:					
IT	.053**	.184*	.155*	.148*	.173
	(.026)	(.102)	(.093)	(.083)	
First lag of IT	−.014	.048	−.098	−.068	.483
	(.020)	(.089)	(.071)	(.109)	
6th grade:					
IT	−.0017	.187**	.117	.144*	.027
	(.016)	(.089)	(.095)	(.081)	
First lag of IT	.023	−.016	−.091	−.052	.161
	(.025)	(.061)	(.064)	(.092)	
Second lag of IT	−.022**	−.056	−.063	−.064	.811
	(.018)	(.056)	(.057)	(.049)	
Individual controls			Yes	Yes	
School controls				Yes	

Note.—Huber-White's robust standard errors are reported in parentheses, corrected for clustering at the school level. The coefficients report the effect of an extra hour of language instruction per week on language test scores. Scores are standardized with a mean of zero and standard deviation of one. IV uses a dummy variable as an instrument, which takes value zero if the specific observation originates from before the policy change and one thereafter. The first and second lags (which are equivalently instrumented, using lagged policy dummies) refer to the effect of the instruction time in language that the students of that group in that particular school experienced 2 and 4 years earlier, respectively. Individual controls are parental education, parental occupation, ethnicity, gender, month of birth, and family structure. School controls are method of instruction, degree of differentiation by ability, amount of homework, frequency of testing, degree of registration of progress (all for both language and math), use of remedial teacher, school finances, number of students in class, and teacher experience. DWH gives the *p*-value of a Durbin-Wu-Hausman test on the endogeneity of each instruction time variable (conducted without controls).
* $p < .10$.
** $p < .05$.
*** $p < .01$.

policy change increased IT for language by around half an hour, its impact is equal to around one-half of the size of the estimates in table 4.

For comparative purposes, we develop an alternative measure of magnitude based on week of testing. There are small differences in the exact week in which schools let their students take these tests, which appear to be random. An additional week of school attendance prior to the test increases scores by 0.15 of a standard deviation in grades 2 and 4 and 0.06 of a standard deviation in grade 6. This implies that the effect of the policy change is equivalent to around 1.25 extra weeks of school in grades 2 and 6 and around 0.5 extra week of school in grade 4. The policy increased language instruction by 7 percent (the average IT is 7.5 hours). As tests are made after around 25–30 weeks of instruction that academic year, a 7 percent increase in school attendance would be equivalent to around 2 ex-

tra weeks of school. The impact of those 2 extra weeks of school attendance would exceed the impact of the policy change (which is equal to around half of the estimates in table 4). The reason can be that the school attendance effect also incorporates attendance in other classes as well as general maturity effects or that the marginal effect from increases in language instruction within the school week is lower than the average effect.

The results from figure 3 and table 4 imply that there is a fade-out of investment productivity in the long run. This cannot be explained by low self-productivity and low complementarity alone. Even when investments are perfect substitutes, later cohorts should have a higher skill level because the absolute amount of investments differs.[20] Investments in IT have an effect on language achievement in the short run, but since they effectively lower the value of later investments, their net long-term effect becomes zero. Earlier studies by Soar (1978) and Kiesling (1984) have also identified that the marginal effect of language instruction on achievement reduces to zero or even becomes negative beyond a certain level. The result implies that either there is an early saturation point for the relation between quantity of language instruction and language achievement or there is fast depreciation in the learning process for these skills. There is not enough exogenous variation in investment to say which scenario applies here.[21] The above does not mean that the quantity of language instruction is irrelevant for the long-term level of language skills. It means that the old levels of language instruction were already high enough to bring the pupils to their "saturated" levels of language achievement. These results depend on the original level of language IT as well as on the levels of other inputs. Results can be very different at a different margin.

B. IQ Scores

Panel A of table 5 shows the effects of language IT on IQ scores. We identify strong positive effects in the short run. These are larger in magnitude than those of language IT on language scores. OLS estimates are again low and they are statistically significant for IT only in grade 4. For the IV model, both the short-term effects and the first lags are statistically significant and large. We do not observe significant effects of grade 2 IT on grade 6 IQ scores, although the coefficient has a positive sign. The lack of statistically significant findings for the long run suggests that these effects fade away also after a few years, albeit slower than for language achievement. More language instruction could enhance the proficiency to solve

[20] And since investment is higher for the youngest cohort in every grade, this cannot be explained by differences in the relative productivity of investments in different periods; see Cunha et al. (2010).

[21] This would require one comparison group in which the investment takes place in an early period but not in a later one. We do not have such a group since investments are made consistently after the policy change.

TABLE 5
Effect of Instruction Time (IT) for Language on Test Scores in Other Subjects

Explanatory Variable	OLS (1)	IV (2)	IV (3)	IV (4)	DWH (5)
	A. IQ Scores				
2nd grade:					
IT	.022	.380***	.358***	.358***	.000
	(.022)	(.099)	(.103)	(.094)	
4th grade:					
IT	.050***	.289***	.293***	.272***	.003
	(.018)	(.069)	(.070)	(.092)	
First lag of IT	−.0069	.225**	.187***	.092	.006
	(.018)	(.062)	(.056)	(.076)	
6th grade:					
IT	.0045	.242**	.256***	.226**	.025
	(.017)	(.101)	(.103)	(.079)	
First lag of IT	.035***	.219***	.201***	.236***	.000
	(.018)	(.078)	(.071)	(.088)	
Second lag of IT	−.024**	.096	.059	.022	.020
	(.012)	(.111)	(.107)	(.093)	
	B. Math Scores				
6th grade:					
IT	.00090	−.107	−.189*	−.165*	.410
	(.030)	(.094)	(.099)	(.095)	
First lag of IT	.020	.0067	−.129	−.103	.678
	(.017)	(.135)	(.180)	(.131)	
Second lag of IT	.025*	.171**	.224**	.177**	.035
	(.014)	(.085)	(.109)	(.087)	
	C. ES Scores				
6th grade:					
IT	−.062***	−.068	−.082	−.093	.386
	(.024)	(.082)	(.080)	(.074)	
First lag of IT	.030	.0055	.0012	−.013	.892
	(.022)	(.148)	(.184)	(.113)	
Second lag of IT	.025*	.159*	.134	.148**	.035
	(.014)	(.087)	(.084)	(.066)	
Individual controls			Yes	Yes	
School controls				Yes	

Note.—Huber-White's robust standard errors are reported in parentheses, corrected for clustering at the school level. The coefficients report the effect of an extra hour of language instruction per week on IQ, math, and environmental studies (ES) test scores. Scores are standardized with a mean of zero and standard deviation of one. Environmental studies is a combination of biology, geography, and history. For an explanation of the IV approach and the control variables included, see table 4. DWH gives the *p*-value of a Durbin-Wu-Hausman test on the endogeneity of each instruction time variable (conducted without controls).
* $p < .10$.
** $p < .05$.
*** $p < .01$.

these IQ tests without making long-term improvements in general intelligence. This finding relates to Jaeggi et al. (2008), who find that working on "unrelated" cognitive tasks can improve performance on subsequent IQ tests. The pattern we find for IQ effects also appears similar to that of several preschool interventions. The Perry Preschool Program led to strong short-

run impacts in IQ of up to 12 points, but these disappear by age 8 (see, e.g., Anderson 2008; Heckman et al. 2010). Similar findings are made for many other preschool and school age interventions (see overviews by Currie [2001], Karoly, Kilburn, and Cannon [2005], and Blau and Currie [2006]). The only exceptions appear to be interventions that already start in infancy (Olds et al. 2004; Heckman, Moon, and Pinto 2010).

C. Achievement in Math and Environmental Studies

Panels B and C of table 5 shows results when we estimate model (2) with achievement in mathematics and environmental studies in grade 6 as outcomes. Only language IT in grade 2 has a statistically significant impact on math achievement in grade 6. We find a similar pattern for achievement in ES. Thus, early investments in language have strong and long-lasting effects on multiple other skills. This is intuitive since students make use of reading and language skills when they are studying other subjects as well. Moreover, math questions in grade 6 tend to be framed in the context of small stories, which probably makes language skills especially valuable there. This effect is present only for IT in early grades, where, presumably, fundamental reading and language skills are shaped. In later grades, language instruction becomes more specific and payoffs to math and ES achievement do not occur. Compared to the pattern in figure 4, the suggested positive effect of IT in grade 4 is very weak in the empirical analysis, while the negative coefficient for IT in grade 6 is relatively strong but not statistically significant.

The estimation for math and environmental studies is based on a smaller sample than that for language and IQ achievement because the results for the high-stakes Cito test are not available for the whole sample. When we estimate the effect of language IT on language achievement for this reduced sample, the results are very similar to those in table 4. The current IT effect is positive and statistically significant (coefficient of 0.157) and the lags are negative and statistically insignificant. Thus, the combined result of only short-term effects of language instruction for language achievement and long-run effects for math achievement is also found when we estimate with a consistent sample across school subjects.

Spillovers can also occur through an intermediate effect on general intelligence. However, the long-term effects for mathematics and environmental studies seem to be too strong to be entirely driven by IQ effects. Furthermore, since we do not have scores for these subjects in multiple grades, we cannot say whether the stronger impact of early language instruction is due to high complementarity over time or a high initial impact. As Cunha et al. (2010) point out, the optimal ratio of early to late investment is positively related to both the degree of complementarity over time and the relative investment productivities in different periods. A value-added model, presented in Section VI.B, can provide more insights into which mechanism is stronger.

D. *Noncognitive Skills*

Section IV.C suggests that the extra time for language instruction largely came at the expense of topics related to the building of citizenship. Such instruction could have benefits for noncognitive skills. The data include factor variables for self-confidence, school enjoyment, and social integration in class based on teacher reports. Results show that the increase in language instruction (and the corresponding decrease in instruction focused on building citizenship) has different impacts in different grades. More language in grade 2 has positive impacts on measures of noncognitive skills in grade 2, while the effect is negative for instruction in grade 4 and statistically insignificant (negative sign) for instruction in grade 6.

E. *Instruction Time in Intermediate Grades*

We observe test scores and levels of IT only for grades 2, 4, and 6. Presumably, the policy also increased language IT in intermediate grades. Some of the estimates can pick up on such variation. For grade 2, the difference between T2 and T3 exists only for more language in grade 2. However, T4 and T5 also (presumably) received more language instruction in grade 1. We can isolate the effect of more language in grade 2 on grade 2 achievement by dropping periods T4 and T5. The estimate decreases from 0.392 to 0.342 of a standard deviation. If we strictly look at the effect of more language in both grades 1 and 2 jointly (dropping T3), the effect increases to 0.410. Hence, the influence of intermediate grades appears small, which is in line with the result that effects from earlier grades fade away quickly.

Not all estimates are affected by potential variation in IT in intermediate grades. The short-term effects in grades 4 and 6 pick up on variation in IT only in that particular grade.[22] The lagged effects can be contaminated, but this does not appear relevant for language achievement as lagged effects are already low and statistically insignificant. However, the long-run effect of language instruction on math achievement picks up on variation in language IT in both grades 2 and 3. The true effect of grade 2 language instruction might therefore be lower than the estimate in table 5 indicates, although the existence of a long-run spillover still holds.

F. *Heterogeneity*

We address potential heterogeneity in treatment effects by estimating model (2) separately for different segments of the achievement distribution. Our method is a variation on the estimation of quantile regressions, first developed by Koenker and Bassett (1978):

$$y_i = x_i'\beta_\tau + u_{\tau i} \quad \text{with } \text{Quant}_\tau(y_i|x_i) = x_i'\beta_\tau,$$

[22] Given that we control for the lagged dummy instruments, the short-term effects in grades 4 and 6 are exclusively based on variation between T2 and T3.

where τ indicates the specific quantile of the distribution. Our approach is similar, but we recode the outcome variable to the rank in the distribution as well. Rank is based on the percentage of people in the complete data set (thus taking all waves) who have a specific number of questions correct. If 65 percent of all pupils have 45 or more questions correct, someone with the exact score of 45 has a rank of 0.35. The coefficients reflect how changes in IT change the rank in the distribution. We split up the ranking into three parts and conduct separate analyses:

$$\begin{aligned}
\text{Score1}_i &= \min(1/3,\ \text{Score}_i),\\
\text{Score2}_i &= \max\left(1/3,\ \min(2/3,\ \text{Score}_i)\right),\\
\text{Score3}_i &= \max(2/3,\ \text{Score}_i).
\end{aligned}$$

The effect for the complete distribution is the sum of the three separate effects. This method is an alternative to the more straightforward approach of conducting split regressions for different groups. One benefit of quantile regression is that it includes all observations in each separate regression. Additionally, it is less vulnerable to outliers.

We estimate model (2) with this alternative outcome variable.[23] The estimates show that the effects of language IT on language achievement are strongest for the low and medium groups in grades 2 and 4 but strongest for the highest group in grade 6.[24] Short-term effects fade away for every subgroup. The treatment effects for IQ scores are strongest for the medium group, while the top group has relatively low short-term effects but has large lagged effects. Hence, the fade-out in effects for IQ scores is lower for the highest achievers. The long-run estimates of language IT on math scores are somewhat stronger for the lower third, but the difference is not large. For environmental studies, the long-run spillover effect is strongest for the middle group and lowest for the top performers.

We also estimate treatment effects separately for boys and girls. The short-run effects of language IT on language achievement are stronger for boys, but they also experience a stronger fade-out. The spillover effect from second-grade language IT to sixth-grade math achievement is stronger for boys, while the effect of fourth-grade IT on sixth-grade math scores is strong in magnitude and close to statistically significant at the 10 percent level for boys. However, none of the differences between boys and girls is statistically significant.

The first stage of our model is too weak to robustly estimate effects by ethnicity. The results for those children with at least one parent born abroad indicate a stronger first lag of language instruction on language scores. However, these are based on an invalid IV model and exhibit very large standard errors.

[23] Heterogeneity results are not portrayed here but are available on request.

[24] This could be due to the nature of the test, which is more diverse and introduces new topics (there are more questions on understanding and interpreting sentences). Better students might be more receptive to instruction in these new topics.

VI. Robustness

The validity of our estimation approach depends on several assumptions. In light of the technology of skill formation presented in Section II, threats to identification exist if there are differences in skill levels (θ_{it}) before the investment takes place or when other investments (I_{it}) have been made during the same time frame.

A. *Differences in Observables*

The comparability of pre- and postpolicy cohorts should be high because we include the same set of schools in every wave. However, student cohorts within a particular school can still be different. We first assess whether differences in observable characteristics between cohorts affect our estimates. A mean comparison of observable characteristics (not shown) reveals no statistically significant differences for ethnicity, parental occupation, month of birth, gender, and whether the child lives in a two-parent household. Parental education, however, is significantly different across time periods. This is part of a general trend of increasing educational attainment over the last few decades. When we add all variables as controls, this has a small negative effect on the short-term effects in grade 6 for both language and mathematics achievement (see col. 3 of tables 4 and 5). The second lag for ES is marginally affected but is not statistically significant anymore ($p = .106$). The long-run effect for math achievement becomes stronger. All these changes stay within one standard error. One could question whether adding controls for parental education gives more reliable estimates. The cross-sectional effect of parental education on test scores is strong also because it picks up on unobserved variables. This relation could be considerably weaker for variation over time, which could lead to "overcontrolling" for increases in parental education. Nonetheless, adding individual-level controls still leaves our conclusions intact and the induced variation is moderate.

B. *Differences in Unobservables*

Pre- and postpolicy cohorts can also differ in characteristics that are not observed but still influence school achievement. An alternative approach would be to estimate a value-added model for an individual panel of students, thereby controlling for baseline performance as a proxy for unobserved ability. The traditional value-added (VA) model specifies current achievement for individual i in cohort c with teacher j in grade g of school s, as follows (Rivkin et al. 2005):

$$\begin{aligned} \text{Score}^{c}_{ijgs} = \beta_0 + \beta_1 \text{Score}^{c}_{ijg-1s} + X^{c}_{ig}\beta_X + T^{c}_{jgs}\beta_T + S^{c}_{gs}\beta_S \\ + f_i + \epsilon^{c}_{ijgs}. \end{aligned} \tag{3}$$

This model controls for prior achievement (Score_{g-1}), family background (X), teacher characteristics (T), school characteristics (S), inherent stu-

dent abilities (f), and a random error term (ϵ). An alternative is to use a gain specification, where prior test scores are subtracted from current test scores on the left-hand side of the equation. The two specifications depend on different assumptions, especially with regard to how the effects of inputs and ability vary with age. Model (3) is more common, but it requires substantial data to assess which method is more appropriate. Todd and Wolpin (2003) provide a more thorough discussion on this. We will present results from both specifications. Kindergarten test scores for the corresponding subject (language or math) are used as a control or to calculate the difference in the gain specification. The kindergarten score in the alternative test (math or language) is also added as a control for student ability. Instruction time in language and in mathematics serve as our main explanatory variables, including their lags when applicable. We include a wide range of individual and school characteristics, including lags for the latter (see the note to table 4 for a list of controls), and also include cohort dummies. We estimate this model separately for each grade.

The main advantage of the VA model over the IV model is that we can now use tests that change over time since we do not have to rely on variation in IT over time to estimate treatment effects. We demean test scores for every wave and grade separately. We can then use baseline test scores from kindergarten for all waves. This also allows us to employ the low-stakes math scores from grades 2, 4, and 6 and the high-stakes Cito test score for language in grade 6 as outcome variables. The main drawback of a VA model is that baseline measures can be susceptible to measurement error. Kindergarten tests might not always reflect the true skills and potential of pupils. The more noise there will be on the baseline test score, the closer the VA estimate will be to the OLS estimate.[25] In addition, this approach requires rich data on various inputs at different stages of the child's life. Especially data on parental inputs are generally scarce. As such, the estimates from the VA model are less reliable than those from a properly instrumented IV model. Still, it is valuable to compare the results of each model, especially when it comes to the observed short- and long-run patterns in the estimates.

Table 6 shows results for both VA specifications. The overall pattern of treatment effects is very similar to that presented in tables 4 and 5. We again find statistically significant effects only for the short-term effects of language instruction on language achievement and for the long-run effect of language instruction on math achievement. The latter finding also provides evidence against the concern that the identified spillover effects from language to math in the IV model could be attributed to the high-stakes test having easier questions in later waves. Interestingly, the impact of language IT in grade 2 on math achievement increases across grades. This suggests that the long-term spillover effect from language to math is

[25] In the extreme case in which the lagged score has no predictive power for current test scores, the model effectively reduces to an OLS model.

TABLE 6
EFFECT OF INSTRUCTION TIME FOR LANGUAGE AND MATH: VALUE-ADDED MODEL

	Language Score				Math Score			
	Grade 2	Grade 4	Grade 6 (LS)	Grade 6 (HS)	Grade 2	Grade 4	Grade 6 (LS)	Grade 6 (HS)
	A. Control							
Language	.042***	.047***	.030*	.039*	.030	.025	−.020	.0090
	(.014)	(.015)	(.017)	(.021)	(.021)	(.016)	(.015)	(.017)
Language first lag		.0085	−.012	−.0015		.029*	−.0060	.010
		(.014)	(.017)	(.019)		(.016)	(.013)	(.016)
Language second lag			−.0064	−.017			.071***	.044**
			(.018)	(.021)			(.017)	(.021)
Math	−.0034	−.011	.027	.024	.0096	.021	.099***	.077*
	(.029)	(.025)	(.037)	(.041)	(.046)	(.027)	(.034)	(.040)
Math first lag		−.0092	−.0054	.0070		.00034	.013	.040
		(.025)	(.025)	(.031)		(.025)	(.022)	(.029)
Math second lag			−.0036	.017			−.047	.010
			(.027)	(.049)			(.029)	(.036)
	B. Difference							
Language	.038**	.049***	.031*	.031*	.0042	.027*	.0035	.0035
	(.015)	(.018)	(.017)	(.018)	(.016)	(.015)	(.017)	(.019)
Language first lag		.0075	−.0032	−.0067		.046***	−.0093	.041**
		(.016)	(.016)	(.022)		(.016)	(.015)	(.018)
Language second lag			−.0087	.0034			.061***	.047***
			(.017)	(.018)			(.020)	(.019)
Math	−.036	−.0062	.024	.030	.030	.029	.102**	.070**
	(.026)	(.028)	(.040)	(.040)	(.032)	(.027)	(.040)	(.033)
Math first lag		−.019	−.0061	.016		−.021	.0029	.013
		(.029)	(.025)	(.029)		(.026)	(.026)	(.028)
Math second lag			−.044	−.022			−.038	.016
			(.029)	(.038)			(.034)	(.036)

Note.—Huber-White's robust standard errors are reported in parentheses, corrected for clustering at the school level. The coefficients report the effect of an extra hour of language and of math instruction per week, including their lags, on standardized test scores of the corresponding school subject, estimated by a value-added model (model [3]). Panel A reports results when we employ kindergarten test scores as a control, and panel B reports results when we use the difference between the scores in kindergarten and the scores in the corresponding grade as an outcome variable. For grade 6, we have data on both a low-stakes (LS) and a high-stakes (HS) test. Tests in lower grades are all low-stakes tests. Scores are standardized to have a mean of zero and a standard deviation of one for every time period and grade separately. We include the same individual and school controls as in tables 4 and 5.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

mainly the result of strong complementarity over time rather than a high initial investment productivity. Table 6 also shows that IT for math provides positive short-term effects on math scores in grade 6, but no long-term estimate is statistically significant. There is no evidence of spillovers from math to language. The grade 6 results are similar for the (low-stakes) PRIMA test and the (high-stakes) Cito test, for both language and math. As the Cito test is specifically designed to discriminate student achievement across the distribution, this result should remove concerns that the lack of long-term effects of language IT on language achievement is due to the grade 6 PRIMA test not being discriminative enough.

Although statistical significance levels are similar, coefficients are substantially smaller in the VA model than in the IV model. They are larger than OLS estimates, but the difference is not very strong.[26] This suggests that using a VA model still leads to substantial underestimation of the effect of IT. This can be due to measurement error on the lagged test scores, which attenuates the estimates in the direction of the OLS coefficients, measurement error on the self-reported explanatory variable, or unobserved heterogeneity.

Table 6 shows effects when we use kindergarten test scores as baseline controls and include all possible lags for IT. We can also assess direct effects in grades 4 and 6 by including only current IT in addition to baseline test scores from 2 years earlier. These results are highly similar to those in table 6: they are slightly higher for grade 4 and slightly lower for grade 6. The VA model also sheds some light on the negative coefficient of language instruction in grade 6 on math achievement in grade 6 in the IV model (table 5). Although the coefficient for language IT is positive in the VA model, we identify a statistically significant and negative interaction between sixth-grade math and sixth-grade language instruction for sixth-grade math achievement.[27] This suggests either that language instruction tends to take time away from subjects that are complementary to math IT in producing math skills or that it reduces what is learned in math classes, for example, by lowering concentration levels. The latter explanation seems more likely given that the increase in language instruction through the policy change came at the expense of time that is presumably less intellectually involving.

C. Mean Reversion

The estimates from the IV model can be biased when the pretreatment cohort simply was a weak cohort. Given that we assess the effects of a policy

[26] The OLS estimates from tables 4 and 5 are based on a different sample: OLS uses a "school panel," while the VA model is based on an individual panel. Additionally, we include wave 6 in the estimation of the VA model. The difference in estimates is more prominent when OLS is applied to the individual panel, including wave 6. Some of the short-term OLS estimates are even negative for that sample. The substantial difference between IV and VA results remains.

[27] The results for the interactions are not shown but are available on request.

change that came in response to poor evaluation results, the results could be explained by an Ashenfelter dip, where performance drops just before treatment and positive effects will be identified by simple regression to the mean.[28] However, the timing of the policy change does not fit the traditional Ashenfelter dip. The poor pupil evaluation results that led to the policy change were obtained from 1988–94, while effective policy changes were made in 1998 only after further deliberation. This also means that the pretreatment cohort falls outside of the evaluation period that the policy followed up on.

It is still possible that the pretreatment cohort coincidentally happened to be a bad draw.[29] First, note that this can explain the positive short-term direct effects only if there was a bad draw of students in wave 2 for all grades. The positive long-run effect for math achievement would require an especially bad wave 4, while the lack of a long-term effect for language achievement would require a strong decreasing trend over time. The results from the VA model already provide evidence against this since they show the same pattern of results when we control for baseline performance for an individual panel of students. Since the two periods before the policy change are not comparable in any of the grade samples, we cannot directly assess whether wave 2 shows a dip in test scores. However, kindergarten test scores are comparable for waves 1 and 2, and they are slightly higher in wave 2 than in wave 1; the difference is 0.06 of a standard deviation for language and 0.03 of a standard deviation for math.

Additionally, we can use kindergarten test scores as controls in the IV estimation for second graders when we limit ourselves to waves 2 and 3 (as these students were in kindergarten in waves 1 and 2). The coefficient for the effect of second-grade language instruction on second-grade language achievement falls from 0.348 to 0.281 when we control for kindergarten test scores, but this difference is not statistically significant. We can conduct a similar exercise for mathematics when we limit ourselves to waves 4 and 5 in grade 6. This allows us to estimate the long-run effect of language instruction on mathematics achievement since one group experiences extra language instruction in grade 2 whereas the other did not. Controlling for kindergarten test scores decreases the coefficient from 0.215 to 0.190, and it remains statistically significant. When we conduct a falsification test that uses kindergarten test scores in math and language as outcomes for the effect of language instruction in grade 2, we obtain coefficients that are low and statistically not significant. Finally, when we restrict the sample to those schools that did not change their hours of language instruction after the policy (we restrict the sample to schools for which the

[28] This is based on findings by Ashenfelter (1978) that there is a preprogram dip in the earnings of participants in job training programs.

[29] As we already compare the same schools over time, this cannot refer to a bad draw of schools, but only to a bad draw of a student cohort within a school.

absolute change is within half an hour), the pattern of mean scores over time is very constant.[30]

D. Trends in Achievement

Rather than a local dip in scores, our estimates could be vulnerable to a trend in achievement over time. Controlling for trends directly is not feasible since all waves differ in the degree of exposure to treatment. Alternatively, we assess how sensitive the results are to incorporating different linear trends (see table 7). When we incorporate substantial negative trends, the lagged effects of language IT on sixth-grade language achievement have a positive sign but are still statistically insignificant, while the lagged effects for math achievement increase.[31] Incorporating a substantial positive trend weakens the effect of early language instruction on math achievement but would also lead to very large negative coefficients for the current effect and the first lag effect and very severe negative lags for the effect of language instruction on language achievement, which both seem implausible. The long-run effect of language instruction on language achievement becomes statistically significant only when we assume a negative linear trend of 0.1 of a standard deviation for every 2 years. Moreover, such a trend would lead to extreme values for the other IT coefficients.

We assess sensitivity to trends for the IQ estimates by taking the "Flynn effect" as a guideline. Flynn (1987) argues that IQ scores increase by 3 points (or 0.2 of a standard deviation) every 10 years. Incorporating such a trend still leads to positive and statistically significant effects for current IT and the first lag of IT. The second lag is reduced to virtually zero. Hence, if the Flynn effect indeed applies here, the fade-out in estimates for IQ would be slightly stronger than in the original estimates.

E. Changes in Quality of Instruction

The previous robustness analyses addressed whether the policy change truly affects student achievement. We now assess whether this change in achievement has operated entirely through changes in language IT or also picks up on differences in school or class quality. Mean comparisons of school quality indicators before and after the policy change show no difference for class size, teacher experience, number of full-time-equivalent teachers, use of remedial teachers, or use of special teachers for pupils whose native language is not Dutch. There is a marginally statistically

[30] The largest difference is an increase in grade 2 from T2 to T3 of 0.07 of a standard deviation. This difference, along with other results discussed in this subsection, suggests that the especially strong estimate for language instruction in grade 2 on language achievement in grade 2 might be slightly overestimated, but this certainly cannot explain the complete effect there.

[31] Alternative data sources such as the Trends in International Mathematics and Science Study and the Progress in International Reading Literacy Study suggest a negative trend in achievement for math and reading over the last decade in the Netherlands; we use this as a guideline here.

TABLE 7
Effect of Instruction Time (IT) for Language on Achievement: Sensitivity to Trends

Explanatory Variable	Original Effect (1)	Trend − (2)	Trend + (3)
Language:			
IT	.187**	.239***[a]	.135
	(.089)	(.091)	(.090)
First lag of IT	−.016	.082[a]	−.113*
	(.061)	(.063)	(.062)
Second lag of IT	−.056	.013[a]	−.125**
	(.056)	(.057)	(.057)
Mathematics:			
IT	−.107	−.072[a]	−.142
	(.094)	(.095)	(.089)
First lag of IT	.0067	.160[a]	−.146
	(.135)	(.146)	(.138)
Second lag of IT	.171**	.244***[a]	.098
	(.085)	(.096)	(.086)
IQ:			
IT	.242**	.290***	.194**[a]
	(.101)	(.112)	(.093)
First lag of IT	.219***	.295***	.143**[a]
	(.078)	(.087)	(.072)
Second lag of IT	.096	.210*	−.018[a]
	(.111)	(.124)	(.102)

Note.—Huber-White's robust standard errors are reported in parentheses, corrected for clustering at the school level. The coefficients report the effect of an extra hour of language instruction per week on the relevant test score (using IV; see model [2]). Column 1 shows the effects as they appear in col. 2 of tables 4 and 5. In col. 2, we extract a negative linear trend in achievement of 0.04 of a standard deviation per 2 years, and col. 3 does the same for a positive linear trend of 0.04 of a standard deviation. On the basis of other data sources, actual trends over this time period are approximated by −0.04 of a standard deviation for math and language and +0.04 for IQ. For an explanation of the IV approach, see table 4.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

[a] The results pertaining to the "appropriate" linear trends.

significant increase in school finances and in the use of end-of-year performance targets. The latter only emphasizes that the new targets for language also effectively reach the average school. The use of midyear targets did not increase. The increase in school finances is consistently present over the last 20 years in the Netherlands. Government reports show that it mainly operates through wage and computerization costs (Centraal Bureau voor de Statistiek 2010). The data show that finances directly aimed at the enhancement of language instruction remain constant. Additionally, the gradual increase in spending cannot explain that students receive a boost in language achievement exactly between time periods 2 and 3 or why there is only a positive effect of language instruction on math achievement for language instruction in grade 2.

Analysis of variables that measure the content of instruction shows only small differences between pre- and postpolicy levels. Some differences can be directly linked to the stronger focus on language instruction, such as increases in the frequency of assigning homework or tests for language. There are no changes for homework and testing for mathematics, nor for the degree of tracking of performance in either math or language. There is a relative increase in time spent on interpreting tables and graphs within language classes. This might have contributed to the positive effect of language on math, although the difference is statistically significant only at the 10 percent level, and controlling for this variable only marginally reduces the estimate. The largest difference we find is for the degree of ability grouping, which is higher after the policy change.[32] The increase is again rather gradual over time and therefore unlikely to drive the specific pattern in scores. Moreover, including these controls in the model does not change our coefficients for language IT. There are no strong differences in the learning materials used for language or math instruction. Hence, we find little qualitative differences in the way in which language (or mathematics) was taught across a wide range of measures. The occurrence of statistically significant differences is very low, especially when we take into account that these are vulnerable to multiple hypothesis testing. Moreover, when we add all school and class input measures to the model (including dummy variables for method of instruction), the changes in the estimates for IT are very minimal (see col. 4 of tables 4 and 5). Most noteworthy is that the direct effect of language instruction on language achievement in grade 6 and the long-term effect of language instruction on ES are now also statistically significant with the inclusion of individual-level controls.

We observed that increases in language IT were the result of decreases in "other time." Although the nature of the policy change suggests that this time was mainly spent on teaching citizenship skills, there might have been reductions in time in which students can work independently, of which a part can be devoted to teaching language skills. If this is the case, our explanatory variable might effectively measure the effect of somewhat less than 1 hour of extra language instruction. There are no precise data on the time that students spent working independently. However, since the long-run estimates for language instruction on language achievement are very low, this is unlikely to be behind the lack of any long-term effect of language instruction on language achievement.

F. Changes in Parental Investment

It is also possible that parents have adjusted their behavior in response to the policy change. We believe that such changes are not likely given the nature and timing of the policy change. News about poor student perfor-

[32] The data show an increase in ability grouping at the expense of both fully homogeneous class instruction and completely individual instruction.

mance already surfaced in 1994; hence any compensating behavior from parents in response to poor education should have occurred earlier. The parental questionnaire does not consistently measure parental involvement across all waves, but the teacher questionnaire asks to what extent teachers agree with the statement "There is a lot of support for reading and general development at the home of this child" on a five-point scale. A mean comparison shows that support is slightly lower after treatment, although this difference is not statistically significant. When we add this variable to the model, this leads to negligible increases in coefficients.

It is still possible that parental investments have changed in some other dimension within the same time frame, as is the case with aspects of the school and classroom that are unobserved. However, we believe that the strong lack of variation over time across the wide range of inputs that we test lends support to the idea that the policy change was rather isolated and that our estimates are not confounded by other changes within the same time frame.

VII. Conclusion

In this paper, we assess the short- and long-term effects of instruction time in language on achievement across various school subjects. We show that changes in the allocation of instruction time (IT) for a specific subject should be analyzed from a multidimensional point of view. The direct effect on achievement in the same subject at the end of the period is one important component, but it is also relevant how the skills for that particular school subject develop over time and whether there are spillover effects toward skills in other subjects. Additionally, one has to take into account that increases in instruction time for one subject take time away from other subjects. An increase in instruction time has net positive effects only when the productive value of the subject that experiences the increase exceeds that of the subject that suffers a decrease in instruction time.

We apply this theoretical concept to a case in which we have exogenous variation in the quantity of language instruction over time through a policy change in Dutch primary education. We find that changes in the quantity of language instruction have both direct and indirect effects on skill acquisition. Direct effects on language skills dissipate quickly, while indirect effects on mathematics and environmental studies are long-lasting.

Hence, in this specific setting, the main value of instruction time in language lies in spillover effects into other subjects. Extensive and early instruction in language is a foundation for skill development in other subjects but has no long-lasting effects for language skills themselves. This dynamic can be compared with good health as a form of human capital. While good health can be very important in fostering and utilizing other skills, the effects of investments in health itself can be very short-term; for example, good physical condition requires continuous physical exercise.

The results from this paper could suggest a similar dynamic for language skills.

The lack of a sustained impact for a quantitative change in language instruction on language skills contrasts with the positive findings of Machin and McNally (2008) for qualitative changes. On the other hand, our results are more in line with findings for US schools that program characteristics seem to affect math scores much more than language scores (Dee and Jacocn 2011; Fryer and Holden 2012; Dobbie and Fryer 2013). However, the baseline circumstances in the United States and the United Kingdom were very different from those in the Dutch setting of this paper, and we do not know to what extent our results generalize to other countries and circumstances. The analysis does not preclude that positive long-run effects of language instruction on language achievement can be present in other situations, nor is there a guarantee that increases in language IT in early grades lead to better math performance later on. Nevertheless, our results indicate that the process of skill formation for math can benefit from having better language skills and that this effect can be sustained even when the direct effect of instruction time for language on language skills does not.

By using an IV approach that exploits a policy change, we correct for the presence of selection bias, which normally plagues estimates of the effect of instruction time on measures of school achievement. We address the possible existence of differences in characteristics between pre- and post-treatment cohorts by controlling for observed characteristics and baseline test scores, as well as by estimating the same effects using a value-added model. The value-added model shows a pattern of results very similar to those for the IV model, although the estimates are smaller in magnitude in the former model. The difference in effect size might result from unobserved heterogeneity in the value-added specification. Measurement error on the baseline score will bias the value-added estimates in the direction of the OLS estimates (Rothstein 2009). The IV estimates are also not affected by controlling for a wide range of school-level characteristics, including variables measuring the quality and content of instruction, nor are they likely to be driven by differences in parental investment. This confirms that the effect of the policy change operated through changes in instruction time rather than qualitative changes made in the same period. Results are also robust to incorporating time trends and variation in the number of included time periods.

Nevertheless, we cannot completely rule out unobserved changes in other inputs within the same time frame as the policy change. Ideally, a field experiment in which one would randomly vary instruction time in different school subjects would be the most optimal approach for estimating the true effects of instruction time, but this type of exogenous variation occurs only very rarely in this context. Such studies also need to include variation in instruction time in subjects other than language to fully cap-

ture the skill technology function and to derive from that what we should teach children during which stages in life.

References

Anderson, Michael. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool and Early Training Projects." *J. American Statis. Assoc.* 103 (484): 1481–95.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Rev. Econ. and Statis.* 6 (1): 47–57.

Bell, Michael L., and Charles W. Davidson. 1976. "Relationships between Pupil-on-Task-Performance and Pupil Achievement." *J. Educ. Res.* 69 (5): 172–76.

Blau, David, and Janet Currie. 2006. "Pre-school, Day Care, and After-School Care: Who's Minding the Kids?" In *Handbook of the Economics of Education*, vol. 2, edited by Eric A. Hanushek and Finis Welch, 1163–1278. Amsterdam: North-Holland.

Centraal Bureau voor de Statistiek. 2010. *Jaarboek onderwijs in cijfers.* Jaarrapport. Hague: Centraal Bureau voor de Statistiek.

Commissie Evaluatie Basisonderwijs. 1994. *Zicht op kwaliteit: Evaluatie van het basisonderwijs: Eindrapport.* De Meern: Inspectie van het basisonderwijs.

Cooley, William W., and Gaea Leinhardt. 1980. "The Instructional Dimensions Study." *Educ. Evaluation and Policy Analysis* 2 (1): 7–25.

Cunha, Flavio, and James J. Heckman. 2007. "The Technology of Skill Formation." *A.E.R.* 97 (2): 31–47.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Currie, Janet. 2001. "Early Childhood Education Programs." *J. Econ. Perspectives* 15 (2): 213–38.

Daniels, Abraham F., and Emil J. Haller. 1981. "Exposure to Instruction, Surplus Time, and Student Achievement: A Local Replication of the Harnischfeger and Wiley Research." *Educ. Admin. Q.* 17 (1): 48–68.

Dee, Thomas S., and Brian A. Jacocn. 2011. "The Impact of No Child Left Behind on Student Achievement." *J. Policy Analysis and Management* 30 (3): 418–46.

Dhuey, Elizabeth, and Justin Smith. 2013. "How School Principals Influence Student Learning." Paper presented at the 2013 Society of Labor Economists conference, Boston, May 4.

Dobbie, Will, and Roland G. Fryer Jr. 2013. "Getting Beneath the Veil of Effective Schools: Evidence from New York City." *American Econ. J.: Appl. Econ.* 5 (4): 26–60.

Flynn, James R. 1987. "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure." *Psychological Bull.* 101:171–91.

Fryer, Roland G., Jr., and Richard T. Holden. 2012. "Multitasking, Learning, and Incentives: A Cautionary Tale." Working Paper no. 17752, NBER, Cambridge, MA.

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educ. Evaluation and Policy Analysis* 19 (2): 141–64.

———. 2003. "The Failure of Input-Based Schooling Policies." *Econ. J.* 113 (485): F64–F98.

Hanushek, Eric A., and Steven Rivkin. 2006. "Teacher Quality." In *Handbook of the Economics of Education*, vol. 2, edited by Eric A. Hanushek and Finis Welch, 1051–78. Amsterdam: North-Holland.

Harnischfeger, Annegret, and David Wiley. 1976. "The Teaching-Learning Process in Elementary Schools: A Synoptic View." *Curriculum Inquiry* 6 (1): 5–43.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251–72.

Heckman, James J., Seong Hyeok Moon, and Rodrigo Pinto. 2010. "The Effects of Early Intervention on Abilities and Social Outcomes: Evidence from the Carolina Abecedarian Study." Manuscript, Univ. Chicago.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Q. Yavitz. 2010. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the High Scope Perry Preschool Program." *Quantitative Econ.* 1 (1): 1–46.

Jacobson, Kerry. 1980. "The Relationship of Individual Student Time Allocation to Reading and Mathematics Achievement." Technical report, Res. and Development Center Individualized Schooling, Univ. Wisconsin–Madison.

Jaeggi, Susanne M., Martin Buschkeuhl, John Jonides, and Walter J. Perrig. 2008. "Improving Fluid Intelligence with Training on Working Memory." *Proc. Nat. Acad. Sci. USA* 105:6829–33.

Karoly, Lynn A., M. Rebecca Kilburn, and Jill S. Cannon. 2005. *Early Childhood Interventions: Proven Results, Future Promise.* Santa Monica, CA: RAND.

Kiesling, Herbert J. 1984. "Assignment Practices and the Relationship of Instructional Time to the Reading Performance of Elementary School Children." *Econ. Educ. Rev.* 3 (4): 341–50.

Kleibergen, Frank, and Richard Paap. 2006. "Generalized Reduced Rank Tests Using the Singular Value Decomposition." *J. Econometrics* 133 (1): 97–126.

Koenker, Roger, and Gilbert Bassett Jr. 1978. "Regression Quantiles." *Econometrica* 46 (1): 33–50.

Machin, Stephen, and Sandra McNally. 2008. "The Literacy Hour." *J. Public Econ.* 92 (5–6): 1441–62.

Olds, David L., et al. 2004. "Effects of Nurse Home-Visiting on Maternal Life Course and Child Development: Age 6 Follow-Up Results of a Randomized Trial." *Pediatrics* 114 (6): 1550–59.

Powell, Marjorie. 1978. "Educational Implications of Current Research on Teaching." *Educ. Forum* 43:27–38.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.

Rothstein, Jesse. 2009. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Educ., Finance and Policy* 4 (4): 537–71.

Sammons, Pamela. 1999. *School Effectiveness: Coming of Age in the Twenty First Century.* Lisse, Neth.: Swets & Zeitlinger.

Scheerens, Jaap. 2000. *Improving School Effectiveness.* Paris: UNESCO, Internat. Inst. Educ. Planning.

Soar, Robert S. 1978. "Setting Variables, Classroom Interaction, and Multiple Pupil Outcomes." Paper presented at the annual meeting of the American Educ. Res. Assoc., Toronto.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.

Stallings, Jane A. 1980. "Allocated Academic Learning Time Revisited, or Beyond Time on Task." *Educ. Researcher* 9 (11): 11–16.

Stallings, Jane A., Margaret C. Needels, and Nicholas J. Stayrook. 1979. "The Teaching of Basic Reading Skills in Secondary Schools, Phase II and Phase III." Technical report, Nat. Inst. Educ., Menlo Park, CA.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg,* edited by James H. Stock and Donald W. K. Andrews. Cambridge: Cambridge Univ. Press.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Econ. J.* 113 (485): F3–F33.

Wang, Jia. 1998. "Opportunity to Learn: The Impacts and Policy Implications." *Educ. Evaluation and Policy Analysis* 20 (3): 137–56.