# New Evidence on the Importance of Instruction Time for Student Achievement on International Assessments*

Jan Bietenbeck†, Matthew Collins‡

August 25, 2021

**Abstract**

We re-examine the importance of instruction time for student achievement on international assessments. We successfully replicate the positive effect of weekly instruction time in the seminal paper by Lavy (*Economic Journal*, 125, F397-F424) in a narrow sense. Extending the analysis to other international assessments, we find effects that are consistently smaller in magnitude. We provide evidence that this discrepancy might be partly due to a different way of measuring instruction time in the data used in the original paper. Our results suggest that differences in instruction time are less important than previously thought for explaining international gaps in student achievement.

*Keywords*: instruction time; student achievement; PISA; TIMSS
*JEL code*: I21

*We thank Kaveh Majlesi, Derek Neal, Therese Nilsson, Luca Repetto, and audiences at Lund University, the autumn 2019 Copenhagen Education Network Workshop, the 2019 PhD Workshop in Education Economics at the University of Stavanger, and the 2020 RGS Doctoral Conference in Economics for helpful comments.

†Lund University and CESifo, DIW Berlin, IZA. Corresponding author. Email: jan.bietenbeck@nek.lu.se. Address: Department of Economics, Lund University, P.O. Box 7080, 220 07 Lund, Sweden. Jan Bietenbeck has no conflict of interest to declare.

‡Lund University. Matthew Collins has no conflict of interest to declare.

## 1 Introduction

Student achievement on international assessments differs widely across countries, and research shows that these achievement gaps are important drivers of cross-country differences in economic growth (Hanushek and Woessmann, 2012). This has spurred interest in the question of what explains international variation in student achievement, with one line of research focusing on the importance of instruction time. In the seminal study in this literature, Lavy (2015) uses student-level data from the Programme for International Student Assessment (PISA) in 2006 to show that weekly instruction time positively affects achievement. Given large variation in weekly instruction time across countries, this suggests that international achievement gaps are partly due to differences in the amount of hours students spend learning in the classroom.

In this paper, we re-examine the importance of instruction time for student achievement on international assessments. We first successfully replicate the results by Lavy (2015) in a narrow sense, using the same student fixed-effects specification and data from PISA 2006 for a sample of OECD countries. We then show that the effect of instruction time is also positive but smaller in data from five further waves of PISA: whereas Lavy (2015) estimates that a one-hour increase in weekly instruction time raises achievement by 0.058 standard deviations (SD), the estimates for the other waves range from 0.014 SD to 0.031 SD. Using data from six waves of the Trends in International Mathematics and Science Study (TIMSS), another international assessment of student competencies, we similarly find smaller effects ranging from 0.015 SD to 0.037 SD. While we are unable to fully explain this discrepancy in results, we provide evidence that the original estimate might be larger partly due to a different way of measuring instruction time in PISA 2006.

In additional analyses, we extend our samples to further countries and show that the effect of instruction time is larger in high-income countries than in low- and middle-income countries, in line with results by Lavy (2015). We also conduct a range of sensitivity checks to gauge whether our estimates are confounded by unobserved factors that the student fixed-effects specification cannot account for. We find no evidence of such bias, but the non-experimental nature of the data does not allow us to completely rule out the influence of confounding unobservables.

Our paper adds to the growing literature on the effect of instruction time on student achievement. Besides the study by Lavy (2015), this research includes important work by Rivkin and Schiman

(2015), who use data from PISA 2009 and two different identification strategies. They estimate impacts of between 0.023 SD and 0.031 SD per weekly hour of instruction. Our results suggest that the smaller magnitude of these estimates could be due to the inclusion of low- and middle-income countries in their sample, or due to the different measurement of instruction time in PISA 2009 compared to PISA 2006. Several other related studies estimate the impact of instruction time on achievement using data for individual countries, including Switzerland, Denmark, and Israel (Cattaneo, Oggenfuss, and Wolter, 2017; Bingley et al., 2018; Lavy, 2020). We contribute to this research by providing comparable international evidence from many different datasets.

## 2 Empirical strategy

PISA is an international repeated cross-sectional study that assesses the competencies of 15-year-old students in math, reading, and science. To estimate the causal effect of instruction time in the resulting individual-level data, Lavy (2015) exploits the fact that each student is observed in three subjects in the following student fixed effects specification:

$$A_{iks} = \beta WeeklyHours_{ks} + \mu_i + \eta_k + \varepsilon_{iks} \tag{1}$$

Here, $i$ denotes students, $k$ denotes subjects (math, reading, science), and $s$ denotes schools. $A_{iks}$ is the achievement of student $i$ in subject $k$. $WeeklyHours_{ks}$ are the weekly hours of instruction received in subject $k$, measured at the school level. $\mu_i$ is a student fixed effect, which controls for all student-level determinants of achievement that do not vary across subjects, such as general academic ability. $\eta_k$ is a subject fixed effect, which controls for any level differences in achievement across subjects. $\varepsilon_{iks}$ is the error term. Lavy (2015) estimates this specification by ordinary least squares and computes standard errors that are robust to clustering at the school level.

The regression in Equation 1 identifies the effect of instruction time from differences between subjects. A causal interpretation of the coefficient of interest $\beta$ relies on the key assumption that there are no other subject-specific determinants of achievement that are correlated with instruction time. We assess the validity of this assumption in Section 4.3 below.

## 3 PISA and TIMSS: background and data

### 3.1 PISA

PISA was first conducted by the OECD in 2000 and has since been repeated every three years. The number of countries participating in the study differs somewhat between waves, but it usually covers more than 50 developed and developing countries. In each wave, PISA draws nationally representative samples of 15-year-old students and assesses them on their math, reading, and science skills using standardized tests. The tests measure students' ability to use their knowledge of the subject to solve real-life problems. Test scores are scaled to have mean 500 and SD 100 across OECD countries participating in PISA 2000. Scores from other countries and later waves are then put onto the same scale, which makes achievement comparable across countries and over time.

Students participating in PISA are asked to complete a questionnaire which, among other things, asks about the weekly amount of school-based instruction time received in each subject. In the 2006 wave of the study, this information was gathered by asking students how much time they typically spend per week attending school lessons in each subject, with possible answers being "no time," "less than 2 hours," "2 or more but less than 4 hours," "4 or more but less than 6 hours," and "6 or more hours." In the other waves, students were instead asked open-ended questions about the number of lessons per week they have in a given subject and how long a typical lesson lasts. Table 1 shows the exact questions used to measure instruction time in each wave.

Our narrow replication uses data from PISA 2006 as in Lavy (2015). For our extension, we use data from five further waves of PISA conducted between 2000 and 2018. We do not analyze data from PISA 2003 because instruction time was measured only in math in that wave, such that we cannot identify its impact using between-subject differences. Following the original paper, we restrict our samples to students who are observed with achievement scores and who answered the questions on instruction time in all subjects, resulting in a balanced panel with three observations per student. The main analysis further restricts attention to a group of 22 OECD developed countries.[1]

The key independent variable in our regressions measures the weekly hours of school-based

[1] The 22 OECD developed countries included in the main sample in Lavy (2015) and our replication are: Australia, Austria, Belgium, Canada, Germany, Denmark, Spain, Finland, France, Greece, Ireland, Iceland, Italy, Japan, Luxembourg, Netherlands, Norway, New Zealand, Portugal, Sweden, Switzerland, United Kingdom.

instruction received in a given subject. In the data from PISA 2006, we follow Lavy (2015) and transform the categorical answers into continuous hours by recoding "no time" to missing, "less than 2 hours a week" to 1 hour, "2 or more but less than 4 hours" to 2.5 hours, "4 or more but less than 6 hours" to 4.5 hours, and "6 or more hours" to 6 hours.[2] In the data from the other PISA waves, we multiply the number of lessons by the number of minutes per lesson and divide by 60. For all waves, we then average instruction time at the school-by-subject level.

Table 1: Questions used to measure school-based instruction time in PISA and TIMSS

| **Study** | **Respondent** | **Question** | **Type** |
|---|---|---|---|
| PISA 2000 | Student | In the last full week you were in school, how many <class periods> did you spend in <subject>? | open-ended |
| | Principal | How many instructional minutes are there in the average single <class period>?[a] | open-ended |
| PISA 2006 | Student | How much time do you typically spend per week studying the following subjects? Regular lessons in <subject> at my school: | categorical[b] |
| PISA 2009 and 2012 | Student | How many <class periods> per week do you typically have for the following subjects? | open-ended |
| | Student | How many minutes, on average, are there in a <class period> for the following subjects? | open-ended |
| PISA 2015 and 2018 | Student | How many <class periods> per week are you typically required to attend for the following subjects? | open-ended |
| | Student | How many minutes, on average, are there in a <class period>?[a] | open-ended |
| TIMSS 1995 and 1999 | Teacher | How many minutes per week do you teach <subject> to your <subject> class? | open-ended |
| TIMSS 2003 and 2007 | Teacher | How many minutes per week do you teach <subject> to the TIMSS class? | open-ended |
| TIMSS 2011 and 2015 | Teacher | In a typical week, how much time to you spend teaching <subject> to the students in this class? | open-ended |

*Notes:* The table gives an overview of the questions used to measure school-based instruction time in each wave of PISA and TIMSS. The term "<class period>" is translated to the locally used term in each country. The term "<subject>" is replaced by "mathematics", "science," or "reading" (only in PISA) for the corresponding subject-specific question. In TIMSS, teachers' answers always refer to the class of students participating in the assessment. [a]This question is not asked separately for each subject. [b]Answer categories for this question: no time, less than 2 hours a week, 2 or more but less than 4 hours a week, 4 or more but less than 6 hours a week, 6 or more hours a week.

The outcome variable in our regressions is the subject-specific test score. As in the original paper, we transform raw scores into z-scores by subtracting 500 and dividing by 100. In this way, we can interpret the estimated effects in terms of standard deviations of the test score distribution among OECD countries participating in the first PISA assessment in 2000.[3]

[2] In his paper, Lavy (2015) writes that he merges the "no time" and "less than 2 hours a week" categories, but the publicly available code on the journal website actually changes "no time" to missing. We choose to follow the code in order to replicate exactly the original estimates, but in practice this makes very little difference.

[3] Both PISA and TIMSS use Item Response Theory to score tests and report scores as a set of five or ten so-called

### 3.2 TIMSS

TIMSS is an international assessment of the math and science knowledge of fourth- and eighth-grade students. The study has been conducted by the International Association for the Evaluation of Educational Achievement every four years since 1995 and usually covers more than 40 countries. In each country, nationally representative samples of students are assessed using standardized tests, which measure students' knowledge of the common part of the math and science curricula of participating countries. Test scores are scaled to have mean 500 and SD 100 across countries in TIMSS 1995, with scores from later waves put onto the same scale. TIMSS also asks all math and science teachers of participating students to complete a questionnaire, which collects information on how many minutes per week they teach their subject to the students' class, among other things. Table 1 shows the exact wording of the questions about instruction time fielded in each wave.

TIMSS shares some key features of PISA, such as the international scope and the focus on more than one subject, which notably allows us to estimate student fixed effects models like in Equation 1. However, there are also important differences between the two assessments. Thus, TIMSS tests curriculum knowledge rather than problem-solving skills and does not cover reading. It also focuses on students in specific grades, who do not correspond exactly to the population of 15-year olds surveyed in PISA (eighth-grade students are about 13.5 years old on average). Moreover, only a subset of the 22 OECD countries examined in Lavy (2015) participated in each TIMSS wave. If the effect of instruction time varies along any of these dimensions, we can expect estimates to differ between the two assessments. By examining data from both PISA and TIMSS, we can gain a more general understanding of the importance of instruction time for student achievement.

Our analysis uses data from six waves of TIMSS conducted between 1995 and 2015. We focus on eighth-grade students and construct our data in a way that closely follows Lavy (2015). Specifically, we restrict the sample to students observed with achievement scores and instruction time in both subjects, and we keep only those countries that are included among the 22 OECD countries described above. To measure instruction time, we first assign each student the total hours received in each subject as reported by her teachers and then compute the school-by-subject average. We measure

plausible values. In our analysis, we follow Lavy (2015) and use the first plausible value for each student in each subject as our outcome. We checked that all of our results are insensitive to choice of plausible value.

achievement using the subject-specific test scores, which we transform into z-scores by subtracting 500 and dividing by 100. The estimated effects are thus scaled in terms of standard deviations of the test score distribution among countries participating in TIMSS 1995.

## 4 Results

### 4.1 Main results

Table 2 presents our main results. Column 1 of Panel A shows the effect of instruction time on student achievement in the PISA 2006 data used by Lavy (2015). The results indicate that a one-hour increase in weekly instruction time raises achievement by 0.058 SD. This estimate is exactly identical to the one reported in the original paper and thus constitutes a successful replication in a narrow sense. Columns 2 to 6 of Panel A show results for five further waves of PISA. The impact of instruction time in these samples is also positive but smaller in magnitude, with estimates ranging from 0.014 SD to 0.031 SD. Panel B reports estimates from six waves of TIMSS, which similarly range from 0.015 SD to 0.037 SD. Thus, effects in TIMSS are comparable to those in PISA (except for PISA 2006) despite the differences in sample and test design. However, note that even among these smaller estimates, the effect of instruction time still varies by a factor of more than two.

### 4.2 Investigating differences in estimates between samples

We now explore why the estimates in Table 2 differ so much between assessments. We focus mostly on the differences between PISA waves and consider two possible explanations: heterogeneous treatment effects and the measurement of instruction time. Heterogeneous treatment effects, for example by student characteristics, could account for the divergent results if samples differed between assessments. We discuss this possibility in detail in Online Appendix B. We show that there are only small differences in student characteristics between PISA samples, and we argue that any change in other, including unobservable, sample characteristics is likely gradual over time. We conclude that heterogeneous treatment effects are unlikely to be the main explanation for the differences in estimates between assessments, and especially for the much larger estimate in PISA 2006.

Another possible explanation for the divergent estimates is that the measurement of instruction time varies across assessments. This possibility is closely linked to the variation in the format of

Table 2: Effect of instruction time on student achievement in OECD developed countries

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| **Panel A: PISA data** | | | | | | |
| | Orig. data: PISA 2006 | PISA 2000 | PISA 2009 | PISA 2012 | PISA 2015 | PISA 2018 |
| Weekly hours | 0.058 | 0.019 | 0.027 | 0.031 | 0.020 | 0.014 |
| | (0.004) | (0.003) | (0.002) | (0.002) | (0.002) | (0.002) |
| # of observations | 460,734 | 65,577 | 493,800 | 327,891 | 420,186 | 342,288 |
| # of students | 153,578 | 21,859 | 164,600 | 109,297 | 140,062 | 114,096 |
| # of schools | 6,577 | 4,352 | 7,176 | 7,774 | 6,204 | 6,070 |
| # of countries | 22 | 21 | 22 | 22 | 22 | 21 |
| Mean weekly hours | 3.4 | 3.8 | 3.7 | 3.7 | 3.8 | 3.9 |
| **Panel B: TIMSS data** | | | | | | |
| | TIMSS 1995 | TIMSS 1999 | TIMSS 2003 | TIMSS 2007 | TIMSS 2011 | TIMSS 2015 |
| Weekly hours | 0.037 | 0.037 | 0.024 | 0.015 | 0.019 | 0.017 |
| | (0.006) | (0.009) | (0.007) | (0.004) | (0.007) | (0.006) |
| # of observations | 83,200 | 43,036 | 46,840 | 41,134 | 48,322 | 81,092 |
| # of students | 41,600 | 21,518 | 23,420 | 20,567 | 24,161 | 40,546 |
| # of schools | 1,770 | 949 | 915 | 804 | 918 | 1,324 |
| # of countries | 16 | 6 | 6 | 5 | 6 | 8 |
| Mean weekly hours | 3.3 | 3.2 | 3.1 | 3.0 | 3.3 | 3.3 |

*Notes:* The table shows estimates of the effect of weekly hours of instruction on student achievement in a sample of 22 OECD developed countries. Panel A shows results based on PISA data. Column 1 reports estimates from the 2006 wave of PISA used in Lavy (2015) and columns 2-6 report estimates from five further waves of PISA. Column 2 covers only 21 countries because data on instruction time are not available for Norway in PISA 2000 and column 6 covers only 21 countries because data for reading achievement are not available for Spain in PISA 2018. Panel B shows estimates based on six waves of TIMSS data. The samples in these regressions include the subset of the 22 OECD countries which participated in the corresponding wave of TIMSS. All specifications in both panels control for individual and subject fixed effects. Standard errors in parentheses are robust to clustering at the school level.

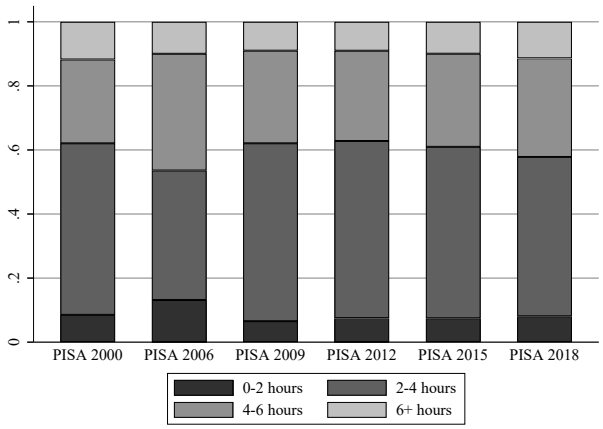
instruction-time questions shown in Table 1. As noted before, PISA 2006 is the only assessment in which students reported instruction time in broad categories of hours. There are also more minor differences in question format among the other assessments, which measure instruction time in minutes. Interestingly, there appears to be an association between question format and effect size: PISA 2006 with its categorical measurement produces by far the largest point estimate; PISA 2009 and PISA 2012 use identical questions and produce similar point estimates, and the same is true for PISA 2015 and PISA 2018 and three pairs of adjacent TIMSS waves.

One way in which question format could affect estimates is by changing the actual answers given by students, teachers, and principals. While we do not observe how the same respondents report instruction time under different question formats, we show that the categorical format used in PISA 2006 is associated with markedly different response patterns compared to the other PISA waves.[4]

[4] The different distribution of student answers in PISA 2006 was also noted by Rivkin and Schiman (2015).

In Figure 1, we graph the share of reported instruction time falling into each of the PISA 2006 categories separately for each wave. Ignoring PISA 2006, about half the students report instruction time in the 2–4-hours category, with only little variation across waves. In contrast, only about 40% of answers fall into this category in PISA 2006. In theory, this difference could reflect an actual change in the distribution of instruction time in 2006. However, Online Appendix Figure A.1 shows that hours distributions before and after 2006 are similar, making this explanation unlikely. Instead, the results suggest that the different question format in PISA 2006 influenced students' answers. Since we do not observe exactly how answers were affected, however, we are unable to establish whether changed response patterns can account for the much larger estimate in that wave.

Figure 1: Distribution of student-reported instruction time across PISA 2006 categories



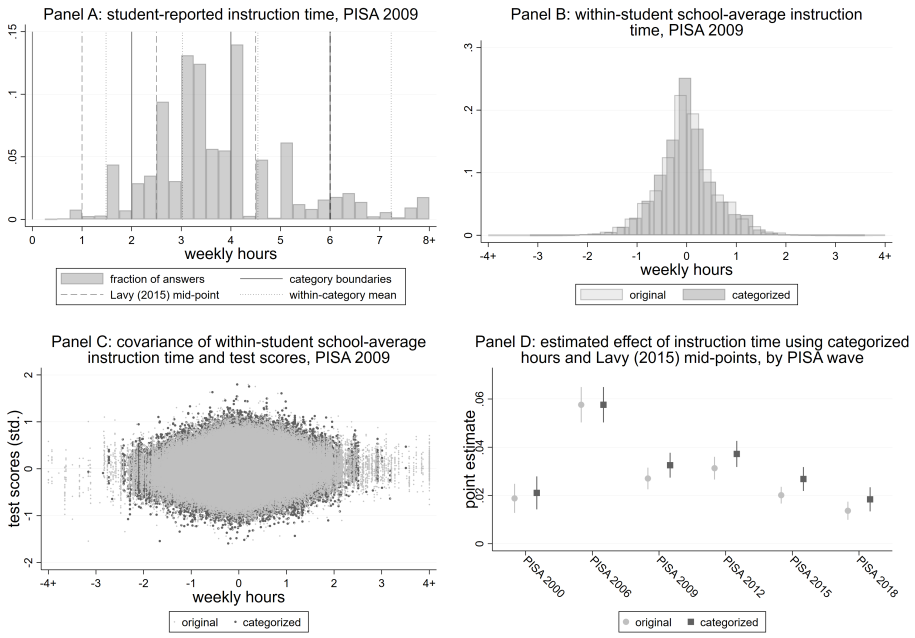

*Notes:* The figure shows the share of student-reported instruction time falling into each of the categories used in the PISA 2006 student questionnaire, separately for each PISA wave.

Another way in which the categorical question format in PISA 2006 could bias the estimate is by introducing an aggregation problem. In particular, the format requires researchers to impute an hours value for each category in order to arrive at a continuous measure, with Lavy (2015) using category mid-points. If the effect of instruction time is linear, aggregation of hours to the within-category mean does not affect the estimate, which implies that this imputation does not matter as long as the mid-point values are equal to the within-category averages. However, if mid-points and averages differ, the estimated effect of hours could be biased either upward or downward.

We illustrate this problem by artificially imposing the answer categories from PISA 2006 onto the hours distribution in PISA 2009. Panel A of Figure 2 shows the distribution of student-reported instruction time together with category boundaries, within-category means, and the Lavy (2015)

mid-points. As can be seen, most mid-points differ substantially from the corresponding within-category mean. Panels B and C show that using these mid-points to aggregate answers leads to a reduction in the variance of within-student school-average instruction time, whereas the covariance between within-student test scores and school-average instruction time does not appear to change much. Put differently, the aggregation reduces the variance of the explanatory variable without markedly decreasing its covariance with the dependent variable, leading to an upward bias in the estimate. Panel D quantifies this bias and shows that the estimated effect increases by 20 percent in PISA 2009 and by between 12 and 34 percent in the other PISA waves if the PISA 2006 categories and Lavy (2015) mid-points are artificially imposed there.

Figure 2: Effects of imposing categorical measurement of instruction time




*Notes:* Panels A, B, and C of the figure show the effects of artificially discretizing instruction time in PISA 2009 by imposing the PISA 2006 answer categories and Lavy (2015) mid-points. Panel A shows the distribution of student-reported instruction time across subjects together with category boundaries, within-category mean hours, and mid-points. Panel B shows the distributions of subject-specific school-average instruction time for the original and categorized variables after residualizing on student and subject fixed effects. Panel C shows a scatter plot of test scores and subject-specific school-average instruction time, both of which have been residualized on student and subject fixed effects, for both variables. Panel D shows how imposing the categorical measurement and Lavy (2015) mid-points affects point estimates in each PISA wave. For details on how the differences between the main estimates and the estimates based on categorized hours in this panel materialize, see Online Appendix Table A.1.

The results in Figure 2 suggest that aggregation using mid-points can account for part of the difference in estimates between PISA 2006 and the other waves. An implication is that using correct

within-category averages should reduce the PISA 2006 estimate. While the hours distribution in that wave is unobserved, we can impute within-category averages using the distributions from the five other PISA waves. Doing this reduces the point estimate from 0.058 SD to 0.047 SD.[5] Taken together, our results suggest that the different measurement of instruction time in PISA 2006 at least partly explains the much larger estimated effect of instruction time in that wave.

### 4.3 Evidence on the validity of the empirical strategy

A potential concern with our results is that they may be biased by subject-specific confounders. For example, if students who are especially gifted in math selected into schools offering relatively more math hours, our estimates would overstate the effect of instruction time. Lavy (2015) conducts a variety of sensitivity checks to test for such bias. For example, he restricts the sample to schools that do not track students by ability, with the intuition being that these schools should be less likely to admit students based on subject-specific academic ability. He finds that his estimates are broadly robust to this and various other sensitivity checks. We successfully replicated his results using the PISA 2006 data and ran comparable regressions for the five other waves of PISA.

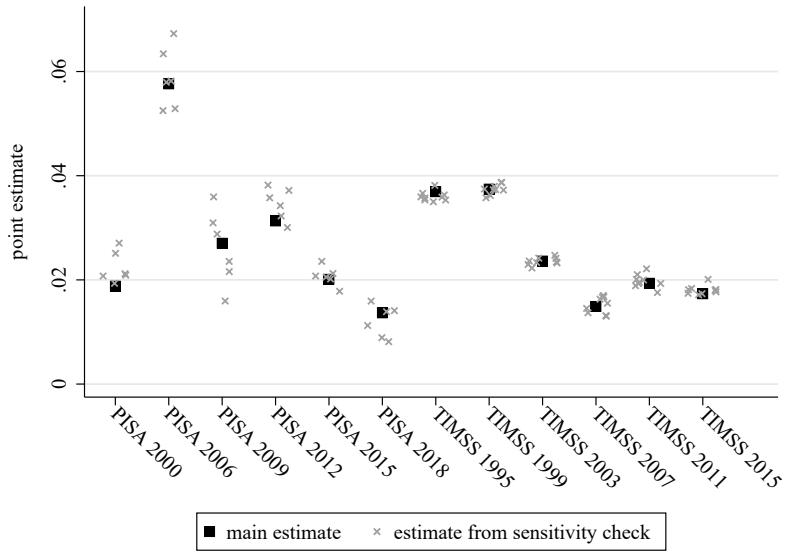
TIMSS collects detailed subject-specific information, which allows us to conduct additional sensitivity checks. Among other things, we observe whether a school offers subject-specific enrichment activities and a proxy for parents' perceived importance of each subject. We included each of these variables as a control in a separate regression. If subject-specific confounders were driving our results, we would expect the estimated effect of hours to change with the inclusion of these controls.

Figure 3 summarizes the results from our sensitivity analysis (full details are presented in Online Appendix C): it plots, separately for each wave of PISA and TIMSS, the main estimate from Table 2 and the estimates from the corresponding sensitivity checks. As can be seen, sensitivity estimates tend to cluster closely around the main estimates, which suggests that subject-specific confounders do not bias our results much. However, we emphasize that we cannot control for all potential confounders, and that the identification assumption is ultimately untestable. This conclusion is in

[5] Note that this imputation cannot account for the changed response pattern and possible changes in the actual distribution of hours in PISA 2006. This could explain why the estimate using the imputed values is still higher than the estimates for the other PISA waves. For the imputation, we used all waves other than 2006 and computed within-category average hours separately by country and subject.

line with that by Rivkin and Schiman (2015), who also caution against potential bias when using within-student between-subject identification.

Figure 3: Estimates from sensitivity checks




*Notes:* The figure shows point estimates of the effect of instruction time on student achievement. The black square reproduces the main estimate from Table 2 for the assessment indicated on the horizontal axis. The grey crosses depict point estimates from the corresponding sensitivity checks. See Online Appendix C for further details.

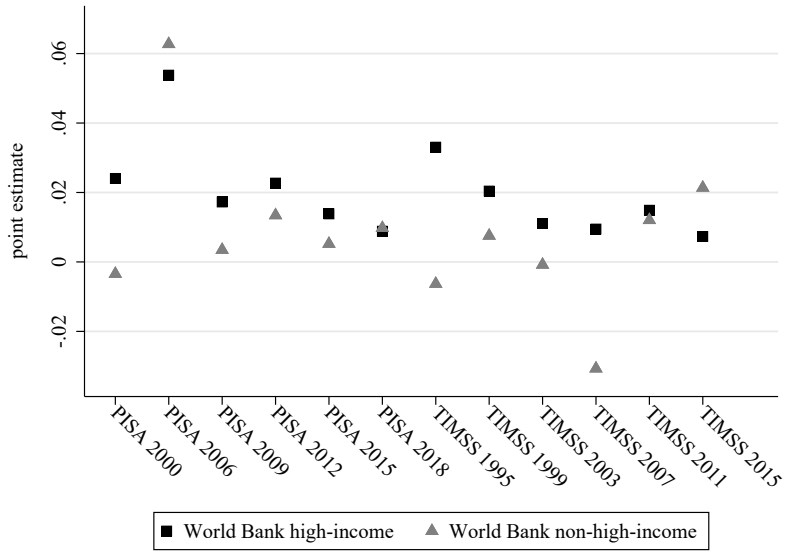
### 4.4 Results for further countries

An interesting question is whether the effect of instruction time differs between developed and developing countries. In the original paper, Lavy (2015) shows that the impact is of similar magnitude to the one found for OECD developed countries in a sample of 14 Eastern European countries, but that it is only about half as large in a sample of 13 developing countries. We successfully replicated these results in a narrow sense and estimated equivalent specifications for the five other PISA waves.

We also estimated the effect separately for high-income and non-high-income economies as classified by the World Bank. This more comprehensive classification, together with the greater coverage across our 12 assessments, means that our results include a total of 59 high-income and 49 non-high-income economies. Figure 4 summarizes our findings (detailed estimation results are shown in Online Appendix Table A.2). It reveals that the effect of instruction time tends to be larger in high-income economies than in non-high-income economies, in line with the conclusion by Lavy (2015). While studying the exact reasons for this difference is beyond the scope of our paper,

one potential explanation is that teachers in developing countries are frequently absent from the classroom, reducing actual instruction time (Chaudhury et al., 2006; Bold et al., 2017).

Figure 4: Estimates for high-income and non-high-income economies




*Notes:* The figure shows point estimates of the effect of instruction time on student achievement separately for high-income and non-high-income economies as classified by the World Bank. For further details and additional estimates, see Online Appendix Table A.2.

## 5 Conclusion

We re-examine the importance of instruction time for student achievement on international assessments. We successfully replicate the estimate of a positive effect of weekly instruction time in the seminal study by Lavy (2015) in a narrow sense. However, when we extend the analysis to data from 11 other international assessments, we find effects that are consistently smaller in magnitude than those reported in the original paper. We show that this discrepancy might be partly due to the different measurement of instruction time in the PISA 2006 data used by Lavy (2015).

Our results suggest that the true effect of instruction time on student achievement is smaller than previously thought. However, some uncertainty about the exact magnitude of the effect remains. One reason is that our smaller estimates still vary by a factor of more than two. Another reason is that the identification strategy relies on the strong assumption that there are no subject-specific confounders. We provide new evidence which suggests that such confounders do not bias our estimates much, but this assumption is ultimately untestable with these non-experimental data.

## References


Bingley, P., E. Heinesen, K.F. Krassel, and N. Kristensen. 2018. "The Timing of Instruction Time: Accumulated Hours, Timing and Pupil Achievement." IZA Discussion Paper No. 11807.

Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson, and W. Wane. 2017. "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa." *Journal of Economic Perspectives* 31:185–204.

Cattaneo, M.A., C. Oggenfuss, and S.C. Wolter. 2017. "The More, the Better? The Impact of Instructional Time on Student Performance." *Education Economics* 25:433–445.

Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan, and F.H. Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20:91–116.

Hanushek, E.A., and L. Woessmann. 2012. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17:267–321.

Lavy, V. 2015. "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries." *The Economic Journal* 125:F397–F424.

—. 2020. "Expanding School Resources and Increasing Time on Task: Effects on Students' Academic and Noncognitive Outcomes." *Journal of the European Economic Association* 18:232–265.

Rivkin, S.G., and J.C. Schiman. 2015. "Instruction Time, Classroom Quality, and Academic Achievement." *The Economic Journal* 125:F425–F448.

# Online appendix

# A Additional results

Online Appendix Table A.1: Effects of imposing categorical measurement of instruction time in PISA waves other than 2006

| Wave | Variable | Residual variance | Residual variance factor | Residual covariance | Residual covariance factor | Point estimate | Point estimate factor |
|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 2000 | original | 0.5165 | 0.7806 | 0.0090 | 0.8750 | 0.0174 | 1.1210 |
| | discretized | 0.4032 | | 0.0079 | | 0.0195 | |
| 2009 | original | 0.3952 | 0.8219 | 0.0104 | 0.9892 | 0.0264 | 1.2036 |
| | discretized | 0.3248 | | 0.0103 | | 0.0318 | |
| 2012 | original | 0.4179 | 0.7632 | 0.0133 | 0.9073 | 0.0318 | 1.1889 |
| | discretized | 0.3189 | | 0.0121 | | 0.0378 | |
| 2015 | original | 0.5541 | 0.5638 | 0.0111 | 0.7517 | 0.0201 | 1.3333 |
| | discretized | 0.3124 | | 0.0084 | | 0.0268 | |
| 2018 | original | 0.4345 | 0.5818 | 0.0059 | 0.7835 | 0.0135 | 1.3467 |
| | discretized | 0.2528 | | 0.0046 | | 0.0182 | |

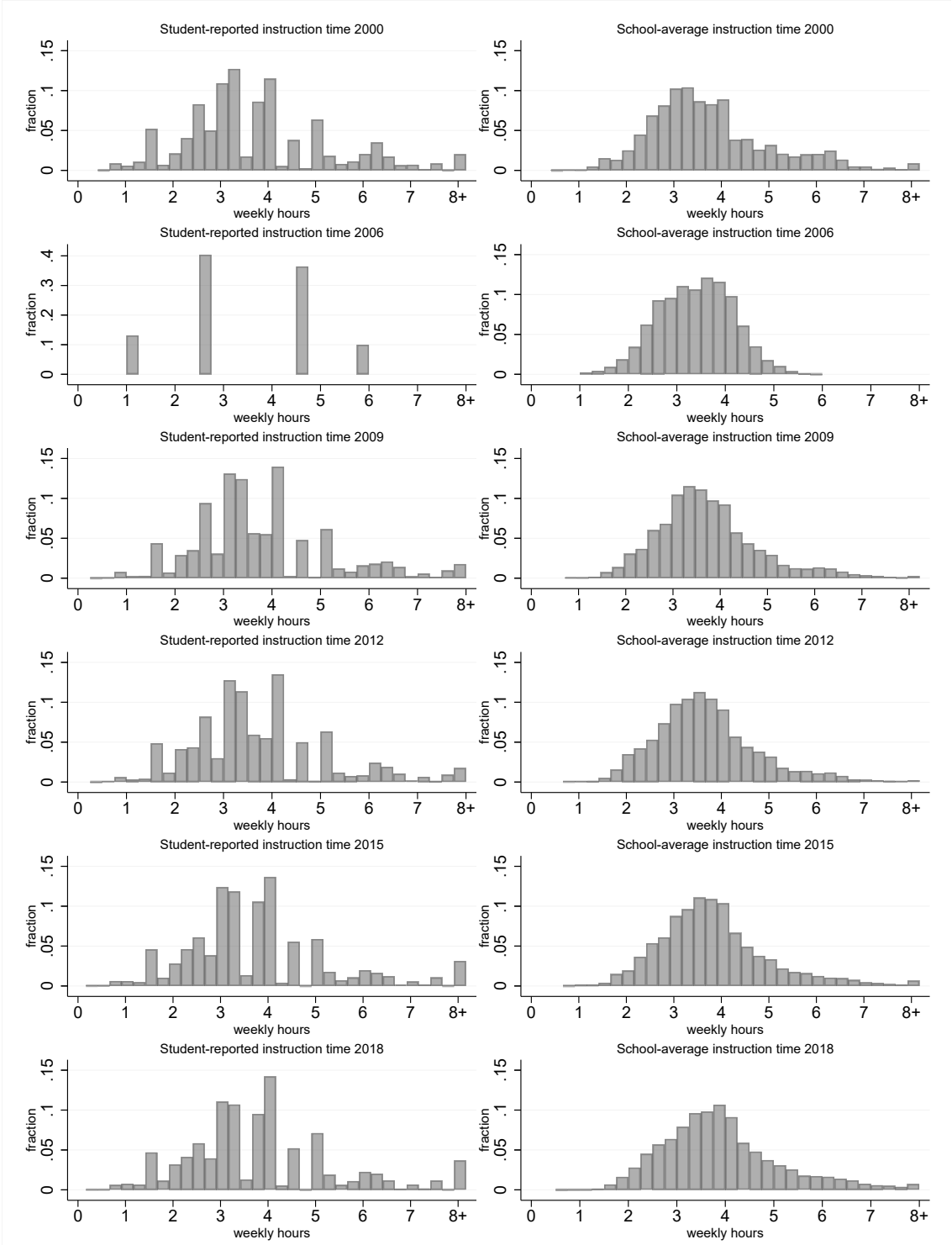
*Notes:* The table shows how artificially discretizing instruction time in PISA waves other than 2006 affects the variance of the explanatory variable (column 3), the covariance between the dependent and explanatory variables (column 5) and, in turn, the point estimate in our main regression (column 7). Instruction time is discretized by imposing the answer categories used in PISA 2006 and the mid-points used in Lavy (2015). Instruction time and test scores are residualized on student and subject fixed effects. Factors reflect the magnitude of the residual variance (column 4), residual covariance (column 6), and point estimate (column 8) when using discretized instruction time, relative to undiscretized instruction time. For further details, see Figure 2 in the main text.

Online Appendix Table A.2: Estimates for different groups of countries

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| **Panel A: PISA data** | | | | | | |
| | Orig. data: PISA 2006 | PISA 2000 | PISA 2009 | PISA 2012 | PISA 2015 | PISA 2018 |
| *A.1: Lavy (2015) 14 Eastern European countries* | | | | | | |
| Weekly hours | 0.061 | 0.023 | 0.004 | 0.018 | 0.005 | 0.009 |
| | (0.006) | (0.007) | (0.004) | (0.004) | (0.004) | (0.003) |
| # of students | 59,005 | 6,416 | 61,147 | 39,062 | 62,932 | 64,984 |
| # of countries | 14 | 7 | 14 | 14 | 12 | 14 |
| *A.2: Lavy (2015) 13 developing countries* | | | | | | |
| Weekly hours | 0.030 | 0.004 | 0.003 | –0.005 | 0.005 | 0.033 |
| | (0.008) | (0.006) | (0.003) | (0.003) | (0.004) | (0.003) |
| # of students | 79,646 | 5,501 | 100,371 | 53,458 | 60,069 | 57,170 |
| # of countries | 13 | 6 | 13 | 11 | 8 | 10 |
| *A.3: World Bank high-income economies* | | | | | | |
| Weekly hours | 0.054 | 0.024 | 0.017 | 0.023 | 0.014 | 0.009 |
| | (0.003) | (0.003) | (0.002) | (0.002) | (0.002) | (0.002) |
| # of students | 227,445 | 28,767 | 273,032 | 169,342 | 256,392 | 241,117 |
| # of countries | 40 | 31 | 47 | 43 | 42 | 46 |
| *A.4: World Bank non-high-income economies* | | | | | | |
| Weekly hours | 0.063 | –0.003 | 0.003 | 0.013 | 0.005 | 0.010 |
| | (0.007) | (0.005) | (0.002) | (0.003) | (0.003) | (0.002) |
| # of students | 91,457 | 9,997 | 144,201 | 76,466 | 79,796 | 138,710 |
| # of countries | 16 | 11 | 26 | 20 | 12 | 26 |
| **Panel B: TIMSS data** | | | | | | |
| | TIMSS 1995 | TIMSS 1999 | TIMSS 2003 | TIMSS 2007 | TIMSS 2011 | TIMSS 2015 |
| *B.1: World Bank high-income economies* | | | | | | |
| Weekly hours | 0.033 | 0.020 | 0.011 | 0.009 | 0.015 | 0.007 |
| | (0.005) | (0.005) | (0.004) | (0.005) | (0.005) | (0.003) |
| # of students | 79,714 | 81,722 | 97,871 | 95,659 | 83,833 | 127,555 |
| # of countries | 34 | 22 | 26 | 25 | 18 | 27 |
| *B.2: World Bank non-high-income economies* | | | | | | |
| Weekly hours | –0.006 | 0.008 | –0.001 | –0.031 | 0.012 | 0.021 |
| | (0.011) | (0.004) | (0.004) | (0.003) | (0.009) | (0.005) |
| # of students | 10,024 | 69,039 | 81,327 | 101,358 | 51,527 | 80,610 |
| # of countries | 4 | 15 | 21 | 25 | 9 | 13 |

*Notes:* The table shows estimates of the effect of weekly hours of instruction on student achievement separately for different groups of countries. Following Lavy (2015), Panel A.1 restricts the sample to 14 Eastern European countries and Panel A.2 restricts the sample to 13 developing countries. The number of countries is lower in some columns because not all countries participated in all rounds of PISA. Panels A.3 and A.4 (based on PISA data) and Panels B.1 and B.2 (based on TIMSS data) show results for high-income and non-high-income economies as classified by the World Bank as of June 2020. The number of countries included in these regressions varies across samples because not all countries participated in all assessments. All specifications in all panels control for individual and subject fixed effects. Standard errors in parentheses are robust to clustering at the school level.

Online Appendix Figure A.1: Distribution of instruction time by PISA wave




*Notes:* The figure shows the distribution of student-reported instruction time (left panel) and school-average instruction time (right panel) separately for each PISA wave.

## B Differences in estimates due to heterogeneous effects?

In Section 4.2, we mention the possibility that estimates differ across waves due to heterogeneous treatment effects. In this Appendix, we discuss this possibility in more detail. Due to the inherent differences in design between PISA and TIMSS, we concentrate on differences in estimates between PISA waves, with a special focus on the PISA 2006 estimate.

One important dimension of heterogeneity in the effect of instruction time is student background. For example, Lavy (2015) shows that the impact of hours is larger for students with an immigrant background and for students with less educated parents. Similarly, Bingley et al. (2018) find that the effect varies by students' gender and socioeconomic status. If student background differed between samples, this heterogeneity could explain the differences in estimated effects. To explore this possibility, panel A of Online Appendix Table B.1 shows means of students' socio-demographic characteristics separately for each PISA wave. All samples are balanced on gender, but immigration status and parental education trend upwards over time. However, these smooth trends cannot account for the much larger estimate in PISA 2006 compared to all other waves.

The effect of instruction time likely also differs by other, unobserved dimensions of student background. Moreover, it varies with school and class characteristics: for example, Rivkin and Schiman (2015) show that the effect differs by classroom quality. While we cannot determine whether the PISA samples are comparable on all possible dimensions of effect heterogeneity, any changes in such characteristics likely follow similarly smooth time trends as the characteristics observed in panel A of Online Appendix Table B.1 and as such cannot account for the much larger estimate in PISA 2006 compared to the other waves.

A related alternative explanation for the differences in estimates is that the distribution of achievement changes across waves: even in the absence of heterogeneous treatment effects, if the standard deviation of achievement was much larger in PISA 2006, this could explain the higher estimate for this wave. However, panel B of Online Appendix Table B.1 shows that means and standard deviations of test scores are broadly similar across waves, making this explanation unlikely.

Finally, non-linearities in the effect of instruction time could be at play if the distribution of hours changed between waves. In Online Appendix Figure A.1, we show that hours distributions in PISA waves other than 2006 are comparable, and we argue in the main text that the true distribution of

instruction time in PISA 2006 likely looks similar. However, due to the very different measurement of instruction time, the observed distribution of hours in that year differs from those in the other years. This means that we unfortunately cannot establish to what extent non-linearities in the effect of instruction time can account for the larger estimate in PISA 2006.

Online Appendix Table B.1: Summary statistics of students' socio-demographic characteristics and achievement by PISA wave

| | Orig. data | | | | | |
|---|---|---|---|---|---|---|
| | PISA 2006 | PISA 2000 | PISA 2009 | PISA 2012 | PISA 2015 | PISA 2018 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| **Panel A: means of socio-demographic characteristics** | | | | | | |
| Female | 0.51 | 0.51 | 0.50 | 0.50 | 0.50 | 0.51 |
| First-generation immigrant | 0.05 | 0.05 | 0.06 | 0.07 | 0.07 | 0.08 |
| Second-generation immigrant | 0.05 | 0.05 | 0.06 | 0.07 | 0.08 | 0.10 |
| Father has college education | 0.24 | –[a] | 0.26 | 0.28 | 0.33 | 0.36 |
| Mother has college education | 0.22 | –[a] | 0.25 | 0.28 | 0.34 | 0.39 |
| **Panel B: mean and standard deviation of achievement** | | | | | | |
| Mean | 513.42 | 521.62 | 509.76 | 513.17 | 509.38 | 510.34 |
| Standard deviation | 93.28 | 96.12 | 92.69 | 90.92 | 92.56 | 93.20 |

*Notes:* The table shows means of students' socio-demographic characteristics (Panel A) and the mean and standard deviation of student achievement (Panel B) separately by PISA wave as indicated in the column headers. Statistics for each wave are computed for the students included in the estimation sample of 22 OECD countries. [a]Data on parental education in PISA 2000 are not directly comparable to data in the other waves because of a change in the PISA student questionnaire after this wave.

## C Details on sensitivity checks

Section 4.3 summarizes results from sensitivity checks that gauge the extent of bias in our estimates due to subject-specific confounders. In this Appendix, we present additional details on these checks.

### C.1 Checks in the PISA data

Our analysis for the PISA data closely follows Lavy (2015) and comprises five sensitivity checks, the results of which are shown in Online Appendix Table C.1. First, we restrict the sample to schools that do not consider students' academic record in the admission process.[6] Intuitively, such schools should be less likely to select students based on subject-specific academic ability, reducing the potential for bias. Panel B presents the corresponding estimates (Panel A reproduces the main estimates from Panel A of Table 2 to facilitate comparison). Second, based on similar reasoning, we restrict the sample to schools that do not consider students' needs or desire for a particular program as a criterion for admission. The results for this check are presented in Panel C.

Third, we restrict the sample to schools that do not practice tracking (in any subject) between or within classes. The intuition is that schools that practice tracking will be more likely to admit students based on subject-specific academic ability. These schools could also place higher-ability students in classes with more instruction time. Panel D shows the results for the sample excluding these schools. Fourth, Panel E presents estimates for the subsample of public schools, for which subject-specific sorting is less of a concern according to Lavy (2015). Finally, we use information on teacher shortages in each subject. For example, schools that are mathematics-oriented might attract more effective math teachers, which could confound the estimates. Panel F presents estimates from regressions which control for an indicator for a lack of qualified teachers in a subject.

Overall, Online Appendix Table C.1 shows that the estimated effect of instruction time is quite similar across the different specifications within a given PISA wave, which suggests that subject-specific confounders do not bias our results. However, one caveat of these checks is that they mostly

[6] Information on factors considered in the admission process was collected in all PISA waves, but the format of the question posed to principals changed somewhat over time. Question formats also changed for some of the other variables used in our analysis, and in a few cases the question was not asked at all. Whenever information is available, we define our variables such that they most closely resemble the original variables used by Lavy (2015).

rely on information that is not subject-specific, and that they therefore might not fully capture the influence of potential subject-specific confounders. As we describe below, the TIMSS data allows us to partially address this concern.

### C.2 Checks in the TIMSS data

In the TIMSS data, we use detailed background information from surveys to identify potential subject-specific confounders related to schools, teachers, and students. Online Appendix Table C.2 presents the results of our sensitivity checks based on these variables. Note that not all variables are available in all waves. Moreover, the format of the underlying survey questions sometimes changes between waves; in these cases, we define variables as consistently as possible across waves.

Starting with school-related confounders, Panel B shows estimates from regressions in which the sample is restricted to schools that do not use students' academic record in the admission process (Panel A reproduces the main estimates from Panel B of Table 2 for convenience). This sensitivity check is equivalent to one of the checks conducted by Lavy (2015). Panel C shows estimates from specifications that add a control for subject-specific tracking by ability. Intuitively, such tracking could influence school choice and could also be related to instruction time, which in turn could lead to bias in our results. Panel D adds a control for whether the school offers subject-specific enrichment activities and Panel E adds a control for subject-specific remedial teaching. Such special teaching activities are likely to attract students with particularly high or low subject-specific ability, and they might also be related to instruction time. The results in Panels B to E show that our estimates from these checks are virtually identical to our main estimates.

Moving on to teacher-related confounders, Panel F adds a control for whether there is a shortage of teachers in a subject at the school, and Panel G adds a control for whether the school has had difficulties filling open teaching positions in a subject. These controls capture a lack of qualified teachers, which could be related to instruction time and also affect student achievement. Building on this same intuition, Panel H shows results from regressions that control for two observable dimensions of teacher quality: education, measured as an indicator for whether the teacher holds an advanced degree, and experience, measured as an indicator for whether the teacher has been teaching for at least five years. Our estimates in panels F to H are robust to these checks.

Finally, we estimate two specifications that add controls for subject-specific confounders related

to students and parents. Panel I uses the fact that in two waves of TIMSS, students were asked to what extent their mother thinks that it is important for them to do well in each subject. This variable proxies for parents' valuation or preferences over subjects and intuitively relates to subject-specific sorting to schools. The specifications in Panel I add this variable as a control to our regressions. In Panel F, we control instead for an indicator for whether a student receives out-of-school extra lessons in a subject. This variable similarly proxies for parents' or students' subject-specific abilities and preferences. The results show that our estimates are robust to both of these sensitivity checks.

Taken together, the estimates in Online Appendix Table C.2 show no indication of bias due to subject-specific confounders related to schools, teachers, or students. However, as we emphasize in the main text, we cannot control for all potential confounders and the key identification assumption in our empirical model is ultimately untestable.

Online Appendix Table C.1: Sensitivity checks in PISA

| | Orig. data | | | | | |
|---|---|---|---|---|---|---|
| | PISA 2006 | PISA 2000 | PISA 2009 | PISA 2012 | PISA 2015 | PISA 2018 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Panel A: main estimates (for comparison)* | | | | | | |
| Weekly hours | 0.058 | 0.019 | 0.027 | 0.031 | 0.020 | 0.014 |
| | (0.004) | (0.003) | (0.002) | (0.002) | (0.002) | (0.002) |
| # of observations | 460,734 | 65,577 | 493,800 | 327,891 | 420,186 | 342,288 |
| *Panel B: academic record not considered for school admission* | | | | | | |
| Weekly hours | 0.060 | 0.017 | 0.025 | 0.034 | 0.022 | 0.013 |
| | (0.005) | (0.004) | (0.003) | (0.004) | (0.003) | (0.003) |
| # of observations | 266,769 | 29,799 | 265,005 | 146,370 | 162,897 | 171,039 |
| *Panel C: students' needs or desire not considered for school admission* | | | | | | |
| Weekly hours | 0.066 | 0.027 | 0.035 | 0.037 | 0.023 | 0.015 |
| | (0.006) | (0.005) | (0.004) | (0.004) | (0.003) | (0.003) |
| # of observations | 171,687 | 22,122 | 182,931 | 117,144 | 124,785 | 138,258 |
| *Panel D: no tracking by ability between or within classes* | | | | | | |
| Weekly hours | 0.052 | | 0.018 | | 0.018 | 0.009 |
| | (0.007) | | (0.004) | | (0.003) | (0.003) |
| # of observations | 160,188 | | 173,958 | | 170,250 | 123,432 |
| *Panel E: public schools* | | | | | | |
| Weekly hours | 0.061 | 0.020 | 0.031 | 0.035 | 0.019 | 0.011 |
| | (0.004) | (0.004) | (0.003) | (0.003) | (0.002) | (0.002) |
| # of observations | 330,492 | 37,899 | 387,117 | 253,281 | 271,068 | 218,034 |
| *Panel F: control for lack of qualified teachers in subject* | | | | | | |
| Weekly hours | 0.058 | | 0.027 | 0.031 | | |
| | (0.004) | | (0.002) | (0.002) | | |
| # of observations | 460,734 | | 493,800 | 327,891 | | |

*Notes:* The table shows estimates of the effect of weekly hours of instruction on student achievement from various sensitivity checks. See text for details on these checks. No estimates are available for some specifications in some waves because the necessary information is not available in those waves. All regressions in all panels control for individual and subject fixed effects. Standard errors in parentheses are robust to clustering at the school level.

Online Appendix Table C.2: Sensitivity checks in TIMSS

| | TIMSS 1995 (1) | TIMSS 1999 (2) | TIMSS 2003 (3) | TIMSS 2007 (4) | TIMSS 2011 (5) | TIMSS 2015 (6) |
|---|---|---|---|---|---|---|
| *Panel A: main estimates (for comparison)* | | | | | | |
| Weekly hours | 0.037 | 0.037 | 0.024 | 0.015 | 0.019 | 0.017 |
| | (0.006) | (0.009) | (0.007) | (0.004) | (0.007) | (0.006) |
| # of observations | 83,200 | 43,036 | 46,840 | 41,134 | 48,322 | 81,092 |
| *Panel B: academic record not considered for school admission* | | | | | | |
| Weekly hours | 0.036 | 0.037 | | | | |
| | (0.007) | (0.009) | | | | |
| # of observations | 76,704 | 37,808 | | | | |
| *Panel C: control for subject-specific tracking by ability* | | | | | | |
| Weekly hours | 0.036 | 0.038 | | 0.015 | | 0.018 |
| | (0.006) | (0.009) | | (0.004) | | (0.006) |
| # of observations | 83,200 | 43,036 | | 41,134 | | 81,092 |
| *Panel D: control for subject-specific enrichment activities* | | | | | | |
| Weekly hours | 0.036 | 0.037 | 0.023 | 0.015 | | |
| | (0.007) | (0.009) | (0.007) | (0.004) | | |
| # of observations | 83,200 | 43,036 | 46,840 | 41,134 | | |
| *Panel E: control for subject-specific remedial teaching* | | | | | | |
| Weekly hours | 0.037 | 0.038 | 0.023 | 0.015 | | |
| | (0.006) | (0.009) | (0.007) | (0.004) | | |
| # of observations | 83,200 | 43,036 | 46,840 | 41,134 | | |
| *Panel F: control for shortage of teachers in subject* | | | | | | |
| Weekly hours | | | | | 0.019 | 0.017 |
| | | | | | (0.007) | (0.006) |
| # of observations | | | | | 48,322 | 81,092 |
| *Panel G: control for difficulty of hiring teachers in subject* | | | | | | |
| Weekly hours | | | 0.024 | 0.015 | 0.019 | 0.018 |
| | | | (0.007) | (0.004) | (0.007) | (0.006) |
| # of observations | | | 46,840 | 41,134 | 48,322 | 81,092 |
| *Panel H: control for experience and education of subject teacher* | | | | | | |
| Weekly hours | 0.037 | 0.036 | 0.024 | 0.015 | 0.020 | 0.017 |
| | (0.006) | (0.009) | (0.007) | (0.004) | (0.007) | (0.006) |
| # of observations | 83,200 | 43,036 | 46,840 | 41,134 | 48,322 | 81,092 |
| *Panel I: control for mother's stated importance of doing well in subject* | | | | | | |
| Weekly hours | 0.036 | 0.038 | | | | |
| | (0.006) | (0.009) | | | | |
| # of observations | 83,200 | 43,036 | | | | |
| *Panel J: control for extra lessons in subject* | | | | | | |
| Weekly hours | 0.037 | 0.036 | 0.024 | | | |
| | (0.006) | (0.009) | (0.007) | | | |
| # of observations | 83,200 | 43,036 | 45,433 | | | |

*Notes:* The table shows estimates of the effect of weekly hours of instruction on student achievement from various sensitivity checks. See text for details on these checks. No estimates are available for some specifications because the necessary information is not available in those waves. When information is available in a wave but the value on a control is missing for an observation, we impute this information at the sample mean and include a dummy indicating missing values in the regression. All regressions in all panels control for individual and subject fixed effects. Standard errors in parentheses are robust to clustering at the school level.