STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
BE551 Degree Project in Economics
Fall 2025

Beyond the Clock: Exploring Instruction Time, Classroom Quality and Student Performance through PISA 2012, 2015 & 2018

Akanksha Sadhu (26081) and Jenica Rinu Devaprasad (26003)

Abstract

This study replicates and extends the work of Rivkin and Schiman (2015) to examine the interaction between instruction time and classroom quality using PISA data from 2012, 2015, and 2018 across 49 countries. We employ fixed effect models to account for unobserved student and school characteristics. Our estimates suggest that while additional instruction time is positively correlated with academic achievement, these returns show signs of diminishing at higher levels. Furthermore, we observe that the benefits of additional time appear to be heterogeneous, depending on the quality of the classroom environment. The productivity of instruction time tends to be higher in classrooms with high student and teacher-related classroom quality, characterized by factors such as lower student disruption and better teacher preparedness. While results for 2015 diverge, likely due to methodological changes in PISA, the findings from 2012 and 2018 imply that policies mandating longer school hours may be less effective if classroom quality is not also addressed.

Keywords: Instruction Time, Classroom Quality, Academic Achievement, PISA, Cross-sectional Analysis
JEL codes: I21, I28, H52, C31

Supervisor: Abhijeet Singh
Date submitted: 8 December, 2025
Date examined: 15 December, 2025
Discussants: Christian Lilja and Stina Stjernqvist
Examiner: Johanna Wallenius

Acknowledgments

We would like to express our deepest gratitude to our supervisor, Associate Professor Abhijeet Singh, for his invaluable guidance, patience, and insightful feedback throughout the process of writing this thesis. His support was instrumental in shaping the direction of this research.

A special note of appreciation goes to Professor Steven Rivkin at the University of Illinois Chicago and Associate Professor Jeffrey Schiman at Georgia Southern University. As this thesis serves as a replication and extension of their influential work, their research provided the essential framework for our analysis. We are grateful for their contribution to the field, which laid the foundation for this study.

We also extend our sincere thanks to Ahmet Girisken, PhD student at the Stockholm School of Economics, and Marleena Tamminen for their helpful advice, time, and willingness to assist us during this project.

We would also like to thank Professor Pushkin Kachroo at the University of Nevada, Las Vegas, for his mentorship and academic support.

Finally, we extend our heartfelt thanks to our family and friends for their unwavering encouragement, understanding, and support throughout our studies.

Artificial Intelligence Statement

In this paper we use generative artificial intelligence (AI) for a two different purposes. We used the AI tools ChatGPT and Gemini to improve the clarity of language in our text through improving readability and grammar. We also used AI to aid us in generating LaTeX code for formatting the output tables that we created through our Stata code. We would like to emphasize that all research ideas, theoretical framing, methodological decisions, data analysis, interpretation of results, and the full intellectual content of the paper was done by us. AI tools did not contribute to the development of arguments or interpretation of findings.

Contents

1 Introduction **1**

2 Background **3**
2.1 PISA Data Explanation and Link to the OECD . . . 3
2.2 External Validity and Societal Relevance . . . 4
2.3 Key Variables . . . 4
2.4 Purpose . . . 5

3 Literature Review **6**
3.1 Instruction Time and Academic Results . . . 7
3.2 Instruction Time, Classroom Quality and Academic Results . . . 7
3.3 Theoretical framework . . . 9
3.4 Evaluation of the OECD and PISA . . . 10
3.5 Final Remarks . . . 12

4 Data **12**
4.1 The PISA Survey . . . 12
4.2 Raw Data . . . 13
4.2.1 Data Cleaning . . . 13
4.2.2 Student PISA Results . . . 13
4.2.3 Instruction Time . . . 14
4.2.4 Student-Related Classroom Quality . . . 15
4.2.5 Teacher-Related Classroom Quality . . . 16
4.2.6 Inconsistencies Across Datasets . . . 18
4.3 Final Data . . . 19
4.3.1 Summary Statistics . . . 20

5 Methodology **21**
5.1 Identification Strategy . . . 21
5.2 Implementation of the Identification Strategy . . . 23
5.2.1 Baseline Empirical Model . . . 23
5.3 Dimensions of Heterogeneity . . . 25
5.4 Discussion of the Identification Strategy . . . 27

6 Results **28**
6.1 Instruction Time Differences Between Subjects . . . 28
6.2 Instruction Time Effects . . . 31
6.3 Heterogeneity by Classroom Environment . . . 36
6.4 Extension to the Study . . . 40

7 Discussion of Results **45**
7.1 The Conditional Productivity of Time . . . 46
7.2 The PISA 2015 Anomaly . . . 46
7.3 Interpreting the Teacher-Quality Extension . . . 48
7.4 Limitations of the Study . . . 49
7.5 Economic Significance and Policy Implications . . . 49

8 Conclusion **50**

A Appendix **56**

1 Introduction

Policy makers face a paradox when creating educational policies: extending the school day to increase instruction time demands budget allocations that could be invested in increasing classroom quality. Is that the right trade-off to make? Although, according to economic theory, investment in schools should continue until diminishing returns emerge, budget restrictions result in what should be an optimisation problem becoming a forced choice between competing necessities. Intuitively, the belief that increased time-on-task leads to better academic outcomes is widespread and seems logical upon first glance. However, increasing instruction time with the aim of increasing time-on-task faces diminishing returns (Rivkin and Schiman, 2015). This means that simply adding hours to the school day does not guarantee proportional gains in learning. Therefore, even though the added investment into increasing school days may increase school results there are perhaps other investments that would increase school results more than proportionally as those factors have not reached diminishing returns yet. The reasons why added instruction time may have diminishing returns could be due to extensive inefficiencies in the schools, through for example, student bullying behaviour, student disruption, teacher absenteeism, or a lack teacher of preparedness. Despite these potential inefficiencies, many countries have recently embraced policies aimed at increasing the length of the school day or mandating more time for core subjects such as mathematics and language arts (Radinger and Boeskens, 2021), even though evidence suggests that the benefits of additional time depend critically on classroom conditions and instructional quality as seen in this paper.

This paper addresses this issue by revisiting the relationship between instruction time, classroom quality, and academic achievement. The broader literature on this topic has extensively studied instruction time and its interaction with quality on academic results, where quality is often defined through structural inputs such as teacher qualifications, years of experience, or the amount of time students spend studying at home. Literature in this area also consists of theories in educational economics. In this study, we utilise the work of Lazear (2001) and Todd and Wolpin (2003) to establish the theoretical basis for what inputs to study, and why and how we look at our classroom quality term. We draw primary inspiration from Rivkin and Schiman (2015), who study student-related classroom quality on the effectiveness of instruction time. We address a gap in the literature by extending the analysis to include teacher-related dimensions of classroom quality and by examining how instruction time interacts with these new variables, as well as with a combined index capturing both student- and teacher-related classroom environments. We argue that the magnitude of any causal link between achievement and instruction time is endogenous to the combined classroom environment.

As a further extension of Rivkin and Schiman (2015), which studies 2009 PISA data, we extend our analysis to include PISA data from 2012, 2015 and 2018. The educational landscape has shifted dramatically since 2009. This makes a replication this

study both necessary and useful. The correlations observed in 2009 may no longer hold due to significant changes to the classroom environment. These could be caused by the integration of information communication technology (ICT) in schools (Chauhan, 2016), the shift from paper-based to computer-based testing in PISA 2015 (OECD, 2017), and the rise of student access to smartphones that negatively impacts the learning environment (Britannica, n.d), the European migrant crisis (European Commission, 2019), and the rise of tutoring and private lessons (Bray, 2021). For example, an increase in the use of smartphones during lesson time could mean that students spend less time-on-task in class. Thus, instruction time increases would have little to no impact on time-on-task, resulting in small gains in academic achievement as students are not attentive in class. This is also a limitation of our classroom quality index as student attention is not measured. Overall, these factors may have altered student attention spans and the nature of disruption in ways that 2009 data could not capture, nor could have ever predicted. We select 2012 as our starting point to provide a direct extension to Rivkin and Schiman (2015). The 2012-2018 window offers a unique opportunity to examine the Disruption Model by Lazear (2001) during a period of relative structural stability before the exogenous shock of the COVID-19 pandemic. This ensures the results between the testing cycles are comparable.

By extending the analysis to the 2012, 2015, and 2018 PISA cycles and incorporating teacher-related factors, such as teacher absenteeism and preparedness, we test whether the complementarities between time and quality remain robust over a decade of rapid technological and structural change. Our extension to this paper allows us to observe whether the interactions between time and quality hold true over almost a decade of educational results, as well as examine whether student-related or teacher-related classroom environment variables have a larger impact on improving the effectiveness of added instruction time.

Our analysis utilises a rich dataset covering 49 countries[1] that participated consistently across the three PISA waves we are examining. By leveraging school-by-grade and school-by-subject fixed effects, we attempt to isolate the effect of instruction time from confounding factors such as non-random allocation of instruction hours or endogenous family responses to student performance. Furthermore, this replication allows us to critically examine inconsistencies and specific changes in the PISA framework over the years. For example, the shift in how instructional minutes are reported.

Our findings indicate that while additional instruction time generally correlates with higher academic achievement, the magnitude of this return is non-linear and significantly heterogeneous on high-quality student and teacher-related classroom environment. Our analysis of the 2012-2018 PISA data reveals a relatively stable allocation of educational resources and student and teacher-related classroom environment across the years. PISA results remained broadly stable between 2003-2018. However, average trajectories

[1] See Table 3 on Page 14 for full list of 49 countries in our sample.

in reading and science across OECD countries started to turn negative in the years leading up to 2018 after a peak performance period between 2009-2012. Furthermore, over the 2012-2022 period, performance in 29 out of 63 participating economies deteriorated in at least two subjects, highlighting the urgency to understand the inputs that drive effective learning (OECD, 2023b). This suggests that the decade long stagnation and decline in academic results measured by PISA is caused by factors within the educational systems or society, rather than a one-time exogenous shock.

Given the prominent role of the OECD and PISA results in shaping global education policy, understanding the dynamics seen in the PISA results is vital for informing education policy and budget allocation. Since the returns to instructional time are dependent on the quality of the classroom environment, policies that simply mandate longer hours without addressing issues like student discipline or teacher preparation are fiscally inefficient. By integrating a focus on student and teacher-related factors that impact the classroom environment, this paper offers a more nuanced view of how educational inputs translate into student performance. This provides actionable insights for policy makers navigating the trade-off between extending instruction time and improving classroom quality.

2 Background

2.1 PISA Data Explanation and Link to the OECD

The Organisation for Economic Co-operation and Development (OECD), an intergovernmental body, conducts the PISA, which evaluates the competencies of 15 year old students in mathematics, science, and language arts on a triennial basis through nationally representative samples of educational institutions and their enrolled students (OECD, 2023a). In this paper, we draw upon data from the 2012, 2015, and 2018 PISA assessment cycles across all nations that participated in all three cycles. Our analysis integrates information from three distinct surveys: the student questionnaire completed by a randomly selected subset of students from each school in the sample, the school questionnaire completed by school administrators, and the educational career questionnaire[2] examining academic trajectories and learning practices. Student performance is quantified through PISA's standardised assessment scores, specifically through their use of plausible values, across mathematics, science, and the language arts. Our primary variables of interest, namely instructional time allocation and the student and teacher-related quality of classroom environments, are derived from responses across these three survey components.

The OECD is recognised for establishing benchmarks and providing policy guidance to governments regarding educational systems. They oversee the administration of

[2]The educational career questionnaire is only used for 2015, this is further explained in Section 2.3

the PISA assessments and publish the raw results as well as reports of the results. The OECD facilitates the collaborative development of assessment instruments, enforces rigorous translation protocols and pilot-testing procedures, and implements standardised sampling methodologies that ensure findings are representative of the 15 year old population within each participating nation and that results are comparable across nations and years, allowing for systematic comparisons. Assessment scores are constructed using contemporary psychometric techniques and standardised on a uniform scale. The OECD provides comprehensive documentation of the testing process and rigorous quality assurance practices. Due to the data's comparability, methodological rigor and that the data is publicly available, PISA results are used by many scholars and policy makers. Since the process and results are completely transparent, the PISA data establishes a robust empirical foundation for examining the relationship between instruction time, classroom quality, and educational outcomes. PISA provides cross-sectional data, not panel data, therefore it does not track the same students or classrooms over time.

2.2 External Validity and Societal Relevance

The 2012-2018 PISA cycles provide comparable measures of instructional time and classroom quality from the student and school questionnaires. PISA results are widely used by education ministries and international organisations to compare and evaluate education systems and educational policy reforms. This means the PISA results are socially relevant and externally valid. Many studies show that students who perform better on PISA at age 15 tend to achieve higher levels of education and better labour market outcomes later in life, even after accounting for socioeconomic background (Meehan et al., 2023). Using PISA data allows us to examine how instructional time and classroom quality relate to academic achievement in a way that is representative within countries, comparable across different education systems and countries, and useful for informing policy on improving student learning and educational systems.

2.3 Key Variables

PISA results provide information on a large range of topics which could directly or indirectly impact school results. The three tested subjects are considered to be the core academic subjects in schooling focused towards higher education. In the 2012 dataset, instructional time is reported separately for each tested subject showing both the minutes per class period and the number of class periods per week. In contrast, the 2015 and 2018 cycles only report an average minutes-per-class-period measure. [3] This change in reporting

[3] Due to this change in reporting, to calculate weekly instructional minutes for each subject in these years we combine average class period length with the corresponding number of class periods per week for each subject.

in 2015 and 2018 is important, as we have no way to know if class period duration in minutes differs across subjects. This is problematic because any subject-specific differences in class length goes unobserved, and we have to assume equal duration of class periods for all subjects. This could potentially introduce measurement error into our weekly instructional hours variable.

In our study we look at the student and the teacher aspects of classroom quality. For student-related classroom quality we create an index that measures the student behaviours that make or do not make the classroom conducive to learning. Similarly, for teacher-related classroom quality we create an index that measure teacher behaviours conducive to learning. The basis for both indexes is a question in the school questionnaire which asks: *"In your school, to what extent is the learning environment of student hindered by the following phenomenon?"* as further described in Section 4.2.1.

PISA measures student performance in language arts, mathematics, and science using a large set of assessment items that no individual student completes in full. Instead, students complete part of each of the surveys and the answers to the rest of the survey are imputed from their answers to the questions they received. PISA employs a rotated test design and estimates student ability through Item Response Theory (IRT), which models performance based on the characteristics of the items students answer and patterns observed in the broader population (Okubo, 2022). An individual student's ability cannot be precisely inferred from a limited subset of items. Therefore, PISA reports 5 plausible values for 2012 and 10 plausible values for academic results 2015 onward. These plausible values represent a range of statistically valid values that the student's true proficiency could plausibly take, given their responses and background data.

The PISA data has across cycle inconsistencies as questions differ slightly from cycle to cycle. The main inconsistencies arise from slight changes in how and what questions are asked, also known as the test framework, and the fact that the major domains rotate with each cycle. The major domain is the main subject being tested in that year and they consist of the core academic subjects mathematics, science and language arts. The questionnaires are revised each cycle, so indexes are not always built from the same items, however they do remain largely similar across years.

2.4 Purpose

The aim of our research is to investigate the causal relationship between instruction time and academic achievement, focusing on how the student and teacher-related classroom environment interacts with instruction time to impact the marginal productivity of instruction time using PISA data from 2012-2018. In order to study this, we examine a sample of 49 countries. Therefore, our main research question is:

1. How does the interaction between instruction time and student-related classroom quality influence student academic achievement across PISA 2012–2018?

   H_0**:** Adding more instruction time has the same impact on test scores, regardless of student-related classroom quality.

   H_1**:** Adding instruction time does not have the same impact on test scores, the impact depends on student-related classroom quality.

   Following the main research question, we have a sub question we also aim to answer in this study:

   (a) How do different dimensions of classroom quality (student-related factors, teacher-related factors, and their combined overall measure) inlfuence the effectiveness of additional instruction time on student achievement?

   H_0**:** Adding more instruction time has the same impact on test scores, regardless of student-related classroom quality.

   H_1**:** Adding instruction time does not have the same impact on test scores, the impact depends on student-related classroom quality.

   H_0**:** Adding more instruction time has the same impact on test scores, regardless of teacher-related classroom quality.

   H_1**:** Adding instruction time does not have the same impact on test scores, the impact depends on teacher-related classroom quality.

   H_0**:** Adding more instruction time has the same impact on test scores, regardless of overall classroom quality.

   H_1**:** Adding instruction time does not have the same impact on test scores, the impact depends on overall classroom quality.

This information is vital for policy makers as it informs them on efficient allocation of resources between extending school hours and improving classroom environment. The results of this study can be used to guide policy development regarding the effective design, targeting, and maintenance of education funding.

3 Literature Review

In this section, we look at the existing literature on instructional time, classroom quality, and its impact on academic results for both PISA and non-PISA data. We find that there are two main strands of studies in this area. Firstly, there are those studies that look only at the impact of instructional time on academic results. Secondly, there are studies that look at

the impact of class time together with classroom quality to determine how the interaction of these two factors impacts academic results. There is also a strand of literature that presents economic models that create a framework for empirical studies. We have included two papers that provides such a framework which we use to inform methodology. With all papers we see an aim of identifying the influence of changing instruction time and/or classroom quality on academic achievement to find what factors create the largest impact.

3.1 Instruction Time and Academic Results

Lavy (2015) examines the impact of instruction time on academic performance. This paper provides one of the earliest large scale cross-country analyses of instructional time using the 2006 PISA data, covering over 50 developed and developing countries. The paper uses within school fixed effects models that compare performance across subjects and isolates the causal effect of instructional time. The results show that additional class time significantly increases academic achievement. However, with substantial heterogeneity across contexts. The paper shows returns are higher in developed countries. This is especially true for schools with stronger accountability and more autonomy in managing their budgets and teachers. The differentiation between developed and developing nations in this paper provides an interesting comparison of the heterogeneity that exists across countries in such an analysis.

Lavy (2020) uses Israel's 2004 funding reform, which unexpectedly changed schools' teaching hour vouchers, as a natural experiment to see whether more resources and instructional time raise achievement. Lavy compares each school before and after the reform with school fixed effects. The instrument used in his analysis is budgets with predicted funding, and he then runs a within-school comparison across subjects to check robustness. The findings are that extra resources mainly got used to lengthen the school week and add hours in core subjects, not to change class size. An added hour per week in mathematics, science, or english increases that subject's score by 0.03-0.05 SD. However, extra Hebrew time has much smaller effects. Therefore, we see that adding time in one subject does not spill over to other subjects. A longer school week has similar sized gains as it channels time into these core subjects. Lavy also measures the impact on non-cognitive outcomes. He sees that more time in class leads to slightly more homework and no decline in school or social satisfaction, and no added violence. Furthermore, time-on-task in core classes causally improves achievement, with modest effect sizes and no behavioural costs, and the mechanism is using extra time well inside classrooms.

3.2 Instruction Time, Classroom Quality and Academic Results

The second strand of studies first measure the basic effect of instruction time on academic outcomes, then go on to measure how the quality of instruction or the classroom envi-

ronment can impact the effectiveness of instruction time and therefore impact academic outcomes.

Looking at a prominent study in this area which builds on the paper by Lavy (2015), Rivkin and Schiman (2015) provide a comprehensive analysis of how instructional time interacts with classroom quality to influence student achievement. They use data from the PISA 2009 cycle across 72 countries and use school-by-subject and school-by-grade fixed effects. The authors find that additional instructional time is positively correlated with achievement in mathematics and language arts. Specifically, an additional hour of weekly instruction time correlated to a 0.03 SD increase in achievement with school-by-grade fixed effects and when using the more conservative school-by-subject fixed effects the effect reduces to 0.023 SD per hour. Furthermore, they find that there are diminishing returns at higher levels of instruction time in any one given subject. However, these diminishing returns only appear at high levels of weekly instruction, around the sixth or seventh hour per week, which is above the 95th percentile of the PISA instruction time distribution. They show that the magnitude of the gains of added instruction time depends strongly on the quality of the classroom environment. For schools in the bottom quartile, the marginal effect of an additional hour per week is 0.017 SD. On the other hand, for schools in the top quartile, the marginal effect rises to 0.027 SD. The highest two quartiles for classroom quality show substantially larger gains from added instruction time than the lowest two. This shows that factors such as student disruptions and weak student–teacher relations, meaningfully reduce the productivity of added instruction time. This highlights that simply increasing instruction time may be insufficient without supportive classroom conditions to ensure that added time is being used for time-on-task activities meaning students benefit from it. Since we find this study to be so meaningful to this area of research, we use it as inspiration for our own study.

Wedel (2021) asks whether adding more instructional time raises test scores and whether that extra time works better with higher classroom quality, measured by teachers' formal qualifications. Using TIMSS 2015 data on 4th-graders in 39 countries, she compares each student's mathematics and science hours and score. The study uses student fixed effects design and aggregates data at the school-by-subject level to limit sorting bias. The findings in this paper are that, on average, an extra hour per week increases achievement by 0.03 SD and slightly improves students' attitude toward the subject. Crucially, she finds that time is more productive with better qualified teachers. This means that if the teacher is specialised in teaching a specific subject the gain rises to 0.04-0.06 SD per hour. However, additional years of experience alone do not amplify the effect. In developing countries, average time effects are near zero, but become positive (0.02-0.03 SD) when the teacher is highly qualified. This highlights strong complementarity between time and teacher qualification. This study shows that more time-on-task can raise achievement, but it pays off most in classrooms with qualified teachers. Hence, simply adding hours without strengthening classroom quality is unlikely to influence scores to a large extent. This is

interesting as it contrasts the paper by Rivkin and Schiman (2015) as it examines a different approach to defining classroom quality, in this case considered as teacher qualification and experience.

Niki (2024) studies Japan's 2002 reform that cut the school week to five days, reducing the annual number of lessons, to see how less instructional time affected 4th-graders' test scores and motivation. This paper uses TIMSS 1995 and 2003 data, where the paper compares each student's mathematics and science time and outcomes. They apply a student fixed effects approach and restrict the empirical strategy to classes where the same teacher taught both subjects. The main finding is that more time leads to better results. There are small but meaningful increases in test scores and in two motivation measures, with effect sizes around 0.03 SD per extra weekly hour of instruction time. However, the study adds that time is more productive in higher-quality classroom settings. They show that gains are consistently higher when teachers have 10+ years of experience. Furthermore, benefits are clearest for students who also study at home. This points to complementarities between time, teacher experience, and learning habits. The study shows that reduction in instruction time reduces both cognitive and non-cognitive outcomes. The study shows that policies should pair added instruction time with hiring experienced teachers and supportive learning environments.

Seidlitz and Zierow (2025) study Germany's rollout of voluntary all-day primary schools to see whether longer days help learning by using a natural experiment. The German federal construction subsidies, worth €4 billion and usable only to build or expand all-day programmes, were used as an instrument for student attendance. They used the NEPS panel data. They find no clear gains on standardised mathematics or german test scores, but grades in German improve and students are more likely to enter the academic track when all-day programmes expand. Furthermore, students also report higher school satisfaction, less bullying, spend more time in school and less time on homework at home. This suggests that the measured benefits of this federal scheme come from the structure and support it provides, rather than added instruction time. However, the effects are uneven across society as non-native speakers and children of single parents benefit less, so the programmes do not reduce inequality in measured outcomes. The study employes multiple quality constraints. For example, including the fact that afternoon time often involves supervision, clubs, and homework help with many staff lacking teaching qualifications. Overall, they find that more time at school can help, especially for engagement and german grades/track placement. However, without high quality instructional components the achievement effects on tests are modest.

3.3 Theoretical framework

In this section, we present two papers that present theories on educational model which we apply to guide our study and methodology. The Disruption Model presented by Lazear

(2001) views classroom instruction time as a public good which is susceptible to crowding out effects. The paper models effective educational output as a function of class size and the probability that an individual student is behaving, and thus he calculates the likelihood of uninterrupted learning. The model shows that the marginal productivity of instruction time is strictly increasing in student behaviour, implying that the optimal class size is endogenous to the classroom environment. Specifically, the paper shows that optimal class sizes are larger when students are better behaved and smaller when disruption is frequent. This framework highlights that the "bad apple" effect, where one disruptive student stops the learning for the entire class, creates a strong complementarity between class size and discipline. This suggests that observed class size effects in empirical data may be muted or biased because schools rationalize class sizes based on the underlying behavioural composition of the student body. Therefore, in our thesis we do not use this model to assess the class size of schools but rather as a motivation for why it is important to measure student disruption when looking at the relationship between instruction time and academic results. Furthermore, we use this study to motivate why we examine diminishing returns to instruction time as this paper pinpoints the non-linear effects of instruction time.

Lazear (2001) explains *why* we study the interaction effect, whereas Todd and Wolpin (2003) explain *how* we study this interaction effect and informs our estimation strategy. The paper explains that cognitive achievement is a cumulative process where current performance depends on the entire history of family and school inputs, as well as the student's innate ability. However, empirical estimation often suffers from missing data on historical inputs and therefore are not able to apply this cumulative methodology. Todd and Wolpin (2003) categorize literature into those who assume past inputs do not matter, *contemporaneous specifications*, and those who assume past scores capture past inputs, *value-added specifications*. These types of models rely on highly restrictive and often untested assumptions to deal with the missing data. Therefore, experimental and non-experimental data estimate fundamentally different parameters. The paper shows that experimental studies, such as this one, measure the total effect. This includes the direct effect of the input plus any behavioural changes (indirect effects). For example, if class sizes drop, parents might reduce their effort at home. The experiment captures the net result of both. Thus, Todd and Wolpin (2001) show that in order to measure the relationship between instruction time and classroom quality it requires a rigorous estimation strategy. Specifically, we need fixed effects to account for the unobserved endowments of student that would otherwise bias a contemporaneous specification.

3.4 Evaluation of the OECD and PISA

Komatsu and Rappleye (2017) explore how the mandatory shift to computer-based testing in the PISA 2015 cycle may have introduced significant inconsistencies in the data. Interestingly, they point out that several high-performing East Asian school systems, specif-

ically Japan, Korea, and Hong Kong, saw their scores drop unexpectedly[4]. The authors argue that this decline was likely not caused by lower student ability, but rather by the students' unfamiliarity with using computers for learning and testing, as these school systems have historically had very low levels of ICT usage in the classroom. This suggests that for these specific countries, the 2015 data (especially for Japan, Korea and Hong Kong) may be influenced by a mode effects, which refers to the negative effects that arise when students find it harder to complete a test on a computer rather than on paper.[5].

Schuelka (2013) investigates the test design and technical reports of TIMSS, PIRLS, and PISA assessments, which are three major international large-scale assessments. The study examined how these organisations define their target population and the language used to justify the exclusion of students with disabilities. For PISA tests the study highlights two aspects. Firstly, PISA allows for school-level exclusions under the label "non-covered areas," which is used to remove special needs schools from the sample. Secondly, PISA allows for nationally defined exclusion categories if priorly agreed on, such as excluding students with dyslexia. PISA sets stricter limits on exclusions, only allowing 2% of the population to be excluded, however the fundamental mechanism of excluding disabled students remains part of the design. Schuelka argues that PISA's definition of academic achievement is one-dimensional because it removes those who do not fit the norm. These exclusions are important to consider when writing a study using PISA data as the dataset inherently has selection bias and is not a true random sample. This means that the results we find may not apply to special needs students, who arguably need instructional support the most. Furthermore, different countries may apply exclusion criteria differently, which can skew country regression results.

Sjøberg (2014) presents a critical analysis of the PISA dataset itself. He examines whether PISA actually measures real life skills as it claims, the methodological reliability of translating questions across cultures and the role of the OECD in using PISA to enforce market-driven educational reforms like standardisation. Sjøberg argues that PISA is a flawed measure of educational quality and often leads to paradoxical results. To make the test statistically fair, the test gives questions relevant to specific national contexts, this means they try to measures a universal curriculum that does not exist in schools. He found that translations into Swedish, Danish, and Norwegian contained errors and clumsy phrasing that changed the meaning of the questions. Furthermore, countries with the highest PISA science scores often have students with the lowest interest and motivation for science. PISA scores are negatively correlated with the use of ICT in schools and active teaching methods like inquiry-based experiments. This means that the PISA test penalizes time spent on non-text-based activities like lab experiments or school excursions.

[4]Japan, Korea, and Hong Kong are among the 49 countries we include in our sample

[5]Further information about this can be read in Section 4.2.6.

3.5 Final Remarks

As seen above, many studies have already been done in the area of instructional time and classroom quality to measure academic performance, both with PISA data and with other international and national datasets. However, we see a gap in studies where both student and teacher-related classroom environment variables are studied for their impact on instructional time and academic results. Furthermore, we believe that measuring this relationship over time is highly beneficial as we get insight into whether the patterns stay stable or evolve across the different PISA cycles. Thus, we aim to fill the gap in research on student and teacher-related classroom quality factors interaction effect with instruction time, to provide a more well-rounded understanding of the impact on classroom quality than previous work in this field.

4 Data

In this section, we present the data which we have used and outline the steps involved in cleaning and preparing our data for analysis.

4.1 The PISA Survey

The PISA survey assesses the skills and knowledge of 15 year old students in mathematics, reading, and science. It is conducted every three years, with the first cycle in 2000 and the most recent cycle in 2022 [6]. PISA participation has more than doubled over the last two decades. While the inaugural cycle in 2000 involved just 32 countries and economies, the most recent 2022 cycle saw 81 participants (Council of Ministers of Education Canada, n.d.).

PISA assesses students aged between 15 years and 3 months and 16 years and 2 months who are enrolled in an educational institution at grade 7 or above. The sampling lists used to select schools and students include all individuals meeting these criteria, regardless of the type of institution or whether they are enrolled full-time or part-time. The aim of the survey is to assess whether students can apply what they learn in mathematics, reading and science outside of the classroom, and to gather information on student attitudes, motivation, collaborative problem solving, global competence and creativity. In addition to this, the PISA student questionnaire also makes an effort to assess the socioeconomic status and family background of students (OECD, 2024).

PISA test scores are reported on standardised scales that allow for meaningful comparisons across countries and over time. Each domain (reading, mathematics, and science) is assigned a scale with a mean of 500 and a standard deviation of 100 among OECD countries during the base year for that subject (OECD, 2024).

[6] The most recent PISA survey was initially scheduled for 2021, but was delayed to 2022 due to the COVID-19 pandemic

4.2 Raw Data

PISA datasets are massive, comprising a high number of variables and participants for every assessment cycle. With three survey years being studied in this paper, and a specific area of study, not all variables were needed to study the effects of instructional time and classroom quality on academic performance. Therefore, in order to make the dataset manageable, we cleaned our dataset for each year as outlined in Section 4.2.1. This left us with a feasible, yet descriptive dataset to perform further analysis.

4.2.1 Data Cleaning

We relied on the publicly available replication package by Rivkin and Schiman (2015) to guide the data cleaning process across all three years. However, we adapted the code that cleaned the data to the variables available for each year as some variables, such as class period minutes, are reported differently in 2015 and 2018. Similarly, we excluded out-of-school observations from the cleaned datasets for 2015 and 2018, as this question had been moved to the optional educational career questionnaire, which only 22 countries administered in 2015[7]. While this still allows us to report results for 2015 for a smaller sample of countries, the question was removed entirely in 2018, making it impossible to report comparable results for that year. We ensured our data had exactly what was needed: instructional time measured in both weekly hours and number of class periods per week, classroom quality indexes (both student-related and teacher-related) built from the same question in the school questionnaire, standardised test scores and a consistent set of relevant variables needed to construct fixed effects for 2012, 2015 and 2018.

4.2.2 Student PISA Results

For each subject, the PISA test assesses reasoning and problem solving skills rather than pure knowledge (OECD, 2017), where the results are communicated through scales. Initially, the mean score for OECD nations for all three subjects was 500, and the standard deviation was 100 (OECD, 2009). This means that two-thirds of students across OECD nations got scores falling in the range of 400-600 points. In later cycles of PISA, the test underwent re-calibration in 2015 to ensure comparability with previous testing cycles. This means that the mean score across OECD countries in recent cycles has experienced minor fluctuations around the original 500 points mark (Robitzsch et al., 2020).

While common school assessments require us to be completely sure of the exact score of what each student receives, large-scale assessments such as PISA do not have the

[7]The 2015 educational career questionnaire was administered in the following 22 countries and economies: Australia, Belgium, China, Bulgaria, Croatia, Denmark, Greece, Hong Kong, Hungary, Iceland, Ireland, Italy, Korea, Latvia, Lithuania, Peru, Poland, Slovak Republic, Slovenia, Spain, Thailand, and the United Kingdom. Of these, two countries were excluded from the final analysis due to a lack of overlap with the primary sample, ensuring consistency across results.

same goal. PISA's goal is to know what the 15 year old population's general knowledge and skills are in the three assessed subjects (Wu, 2005). Therefore, it is more important to reduce error in making inferences about the target population, rather than the individual student (OECD, 2009). Because the PISA assessment design involves students answering only a subset of items, it is not possible to determine a single, precise score for each individual. Instead, PISA estimates a posterior distribution of proficiency. Consequently, the results are reported as plausible values, which are random draws from this distribution that represent the range of scores a student might reasonably have achieved (OECD, n.d.). For 2012, we have 5 plausible values for each student, however from 2015, 10 plausible values are provided per student per subject. The aim with this method is to reduce measurement bias for the population (Okubo, 2022).

The ideal method to deal with plausible values in the PISA results is to estimate regressions using plausible values we must estimate separate regressions with each of the plausible values and then average across the resulting estimates. However, we choose the first plausible value and estimate results based on this. This is because with a large sample size the estimates produced from running separate regression are largely similar regardless of which plausible value is used. Furthermore, similar to Rivkin and Schiman (2015), the estimates were insensitive to the choice of plausible value, thus this choice does not bias the results. We have removed observations with missing plausible values, which resulted in a negligible drop in the sample size.

4.2.3 Instruction Time

Instruction time is measured through two components for each year which we derive from the *minutes per class period for each subject* and *number of class periods a week for each subject* variables in the PISA datasets:

1. Weekly minutes of class time for each subject[8]
2. Weekly number of class periods for each subject

These measures are provided separately for mathematics, language arts, and science. To address implausible values, we follow thresholds from Rivkin and Schiman (2015). Responses reporting more than 10 class periods per week or more than 120 minutes per class are set to missing. The sample is also limited to 9th- and 10th-grade students to ensure better comparability across countries. A key constraint of the PISA design is that students are randomly assigned different test booklets, meaning not every student is assessed in all domains (Perkins and Shiel, 2016). To ensure a consistent within-student comparison for our fixed effects model, we restrict our sample to those students who received booklets containing both mathematics and language arts questions. Since our analysis relies

[8]From this, we derive weekly hours of class time for each subject as well.

on information from both mathematics and language arts, we exclude students who do not take both subjects components, as well as those who take the one-hour version of the test designed for students with special needs. We exclude science test scores from our main sample because we assume there are likley to be spillover effects between mathematics instruction time and science test scores, as mathematics is a fundamental tool used in science[9] (Rivkin and Schiman, 2015). as expected, those with missing instruction time responses are also excluded from our sample. After restricting our sample based on the explanation above, and dropping observations that cannot be matched between the school questionnaire and student questionnaire, we present below our approximate sample sizes for each year:

- **2012**: The main sample for this year comprises 257,016 observations from 124,456 9th and 10th-grade students across 13,263 schools.[10]
- **2015**: The main sample for this year comprises 513,778 observations from 257,432 9th and 10th-grade students across 10,801 schools.
- **2018**: The main sample for this year comprises 382,658 observations from 192,292 9th and 10th-grade students across 12,290 schools.

4.2.4 Student-Related Classroom Quality

Student-related classroom quality is a composite measure that we create using a question in the school questionnaire that provides information on the school environment. The exact question in the school questionnaire is asked as follows: *In your school, to what extent is the learning of students hindered by the following phenomena?* (OECD, 2012). Table 1 reports the distribution of 'A lot' responses[11] to this question, meaning it displays how many schools felt that learning was hindered the most by these specific student-related issues. This question asks about multiple school, teacher and student-related factors that would impact the learning environment in a classroom. Across the three testing cycles the measures used to build student-related classroom quality are: *student truancy, students skipping classes, students lacking respect for teachers, students using drugs and/or alcohol, and students bullying and intimidating others*. We choose questions that are con-

[9] Science scores are included in the appendix analysis to increase statistical precision by expanding the sample size, serving as a robustness check that confirms the main findings persist even when introducing the bias caused by mathematics-science spillovers.

[10] Consistent with the approach in Rivkin and Schiman (2015), we reshape the data from the student level to the student-by-subject level to exploit within-student or within-school variation. To ensure a balanced panel, the sample is restricted to students with valid test scores and instruction time data for both mathematics and language arts. Consequently, each student contributes exactly two distinct observations to the dataset, resulting in a total observation count ($N = 257,016$) that is larger than the number of unique students ($N = 124,456$)

[11] School administrators assessed the extent to which student learning is hindered by specific behaviours on a four-point scale: 'Not at all', 'Very little', 'To some extent', and 'A lot'. The response 'A lot' signifies the highest level of hindrance, indicating a severe disruption to the learning environment.

sistent between all three testing cycles and therefore provide a consistent and comparable measure between the years.

Table 1: **Share of 'A lot' Responses to Student-Related Classroom Quality Measures for PISA 2012-2018**

'In your school, to what extent is the learning of students hindered by the following phenomenon?'			
Year	2012	2015	2018
Student truancy	0.19	0.17	0.18
Students skip classes	0.19	0.19	0.18
Students lack respect	0.28	0.24	0.25
Student alcohol/drug use	0.59	0.54	0.53
Student bullying	0.35	0.31	0.32

Notes. Values represent the proportion of school representatives responding "A lot" to the question: "In your school, to what extent is the learning of students hindered by the following phenomena?" This question is consistent across PISA 2012–2018. The full distribution of responses is available in Appendix Table A1 on Page 44. (Note: For the analysis, responses are reverse-coded so higher values indicate a higher quality environment) **Data source:** The OECD and PISA

To create the student-related classroom quality index, we perform a factor analysis based on the school administrator responses to the list of factors as shown in Table 1. All answers are changed so that a higher value on the index of student-related classroom quality reflects a classroom environment that creates more effective learning for students. Consistently across the years, *student alcohol and/or drug use* was identified as causing the most hindrance in the learning of students, with 53%-59% of schools in the sample reporting that this phenomenon hinders student learning 'A lot'. The student-related classroom quality index shows the isolated impact of the peer environment and student behaviour on learning. The index acts as an examination of the Disruption Model (Lazear, 2001) and allows us to differentiate between total class time and time-on-task. We start by measuring the results on the student-related classroom quality index as this is a direct replication of the method applied by Rivkin and Schiman (2015) in their paper. In the paper, student-related variables are used as proxies for the overall classroom environment as they argue that student-teacher relations and disruption reflect the teachers ability to manage the classroom and influence the environment. However, in order to maintain comparability across the years we cannot include all the factors that Rivkin and Schiman (2015) included in their study as the same questions are not asked across cycles. The most notable change is the drop of the *student-teacher relationship* variable which is not present in 2015 or 2018, which is an important factor for this index to be a proxy for overall classroom quality. This is one of the many reasons why we deem it necessary to also measure the teacher-related classroom quality factors in our study in order to be able to draw conclusions about the overall classroom environment.

4.2.5 Teacher-Related Classroom Quality

While student behaviour significantly influences the classroom environment, teacher-related behaviours are equally important in creating an environment conducive to learning, as out-

lined in Section 3.2. Hence, we use the same question mentioned in Section 4.2.4, and form our teacher-related classroom quality variable using the following factors: *teachers being too strict with students, teachers not meeting student needs, teacher absenteeism, staff resisting change,* and *teachers not being well-prepared*. In specific, *staff resiting change* is important in determining classroom quality as a refusal to adapt to trends like digitalisation in the classroom may cause stagnation, and the prioritisation of comfort over better student outcomes. Likewise, *teacher absenteeism* is a proxy for instructional consistency which is important for students to learn effectively. The rest of the factors we chose are important in a similar manner. Additionally, we chose these questions because they are consistent between all three testing cycles and therefore provide a consistent and comparable measure between the years. However, it is important to clarify that we are not exploring teaching quality, but rather how specific teacher-related factors, such as teacher absenteeism and teachers not being well-prepared, affect the classroom's learning atmosphere.

Table 2: **Share of 'A lot' Responses to Teacher-Related Classroom Quality Measures for PISA 2012-2018**

'In your school, to what extent is the learning of students hindered by the following phenomenon?'			
Year	2012	2015	2018
Teachers too strict with students	0.33	0.32	0.31
Teachers not meeting students needs	0.27	0.29	0.22
Teacher absenteeism	0.39	0.37	0.37
Staff resisting change	0.29	0.29	0.29
Teachers not well-prepared	0.40	0.35	0.34

Notes. Values represent the proportion of school representatives responding "A lot" to the question: "In your school, to what extent is the learning of students hindered by the following phenomena?" This question is consistent across PISA 2012–2018. The full distribution of responses is available in Appendix Table A2 on 44-45. (Note: For the analysis, responses are reverse-coded so higher values indicate a higher quality environment) **Data source:** The OECD and PISA

Table 2 reports the distribution of 'A lot' responses to this question, meaning it displays how many schools felt that learning was hindered the most by these specific teacher-related issues for each year. We create the teacher-related classroom quality index in the same way as our student-related classroom quality index. This means that we look at the teacher-related factors in the question listed in Table 2. It can be observed that *teachers not well-prepared* was identified as the highest hindrance to student learning in 2012. Conversely, in 2015 and 2018, *teacher absenteeism* emerged as the major factor that hinders the classroom learning environment. We take all factors that are common across all three assessed PISA cycles. The teacher-related factors allow us to measure and compare the effect of student behaviour and teacher behaviour on the overall classroom environment, as we acknowledge that both students and teachers may have different amounts of influence on the classroom environment and may influence it in different ways. Furthermore, we see the method of measuring student and teacher effects and combining them to measure overall classroom quality effect as a more robust and precise method of measuring the impact of the overall classroom environment as there are less assumptions made about

how the classroom learning environment is influenced by teachers or students. This makes our methodology more robust as less assumptions in the inputs means that we have less measurement bias in our independent variables, which will lead to less bias in our results (Todd and Wolpin, 2003).

4.2.6 Inconsistencies Across Datasets

To start broadly, there are many technical inconsistencies across the PISA cycles we study. For instance, three domains (mathematics, science and language arts) are tested every cycle, but the major domain tested rotates triennially. The PISA 2012 study focused on mathematics, the PISA 2015 focused on science, and language arts was the major domain for PISA 2018. The major domains receive about double the time of testing than the minor domains, which means that the rotation of the major domain makes the dataset not perfectly identical from cycle-to-cycle (National Center for Education Statistics, n.d.).

Likewise, PISA 2015 was the first time a computer-based version of the survey was offered to schools on an optional basis. Countries that wanted to opt-out of the digital format were allowed to do so. In fact, in 2015, 15 countries[12] chose to opt-out of the computer-based test. Our sample of 49 countries for 2015 consists exclusively of computer-based test scores. Although this trend continued in 2018, it is safe to assume that this change would have a stronger effect, if any, between 2012 and 2015. This is because survey organisers seemed to have taken the challenges faced by countries in the PISA 2015 cycle into serious consideration, thereby developing a strategy to reduce disruption. One of the ways they addressed this is by implementing multi-stage adaptive testing in 2018 which is detailed further in the next paragraph. Likewise, 2018 was the first time PISA launched a 'Student-Delivery-System manual'[13]. The main purpose of this manual is to ensure standardised testing conditions across all schools (OECD, 2016). Moreover, research into the 2015 cycle shows that mode effects likely influenced the comparability of scores from the 2012 cycle, with mean scores from the computer-based test lower than the mean scores of the paper-based test (Jerrim et al., 2018). In addition, it is important to note that Japan, Korea, and Hong Kong, which are identified by Komatsu and Rappleye (2017) as have been the most adversely effected from the shift to digital testing in 2015 the most, are also included in our sample.

As mentioned before, PISA 2018 introduces a new dimension of methodological complexity. Multi-adaptive testing was introduced in 2018 for language arts scores, which meant that a three-stage process was introduced to adapt the difficulty of questions to each student's ability. This made results more precise as it factored in student's ability. For the

[12]The 15 countries that opted-out in 2015 are as follows: Albania, Algeria, Argentina, Georgia, Indonesia, Jordan, Kazakhstan, Kosovo, Lebanon, Macedonia, Malta, Moldova, Romania, Trinidad and Tobago, and Vietnam (Jerrim et al., 2018).

[13]The PISA 2018 Student Delivery System (SDS) is a self-contained software suite designed to administer computer-based assessments and questionnaires through a secure, locked-down browser interface, ensuring standardised testing conditions across different school hardware. See PISA Student Delivery System Manual for more information.

first time, an equitable testing structure was implemented (National Center for Education Statistics, 2019).

Now, coming back to the inconsistencies between the datasets in specific, there are two main differences that we would like to highlight. Firstly, the way minutes per class period reporting changes between 2012 and 2015. While PISA 2012 clearly reports minutes per class period for each of the three subjects, PISA 2015 and PISA 2018 both report average minutes per class period, not per subject. Therefore, when we derive weekly minutes of class for each subject, we obtain a less precise estimate for all three subjects for the 2015 and 2018 cycles. However, we acknowledge that this introduces a layer of measurement error - by relying on a generic average, we lose the ability to detect specific variations in instructional time allocation between mathematics and science. That being said, because the mean duration of a standard class session remain stable across all studied survey cycles, consistently ranging between 51-53 minutes (0.85–0.90 hours)[14], it is safe to assume that our analysis stays consistent across the three PISA cycles.

Although there are some issues to take into account with inconsistencies in the PISA data across the years, these issues do not invalidate the dataset as we adjust for these factors in our methodology and ensure all variables included in our analysis are comparable throughout the years. In this thesis, we (i) use only the items common to 2012, 2015, and 2018, or ones that have been made comparable; (ii) use OECD plausible values correctly; (iii) include school-by-grade and school-by-subject fixed effects to absorb framework shifts; and (iv) interpret the classroom quality indexes we build as proxies, acknowledging that stronger causal claims would need linked teacher qualification data or observational measures not available in PISA.

4.3 Final Data

To ensure a direct comparison across PISA cycles, we include the 49 countries common to all years. This still provides an extensive dataset that allows us to thoroughly investigate our research question. However, it is important to note that each PISA cycle tests a different group of students, as it is a cross-sectional database that does not track individuals over time. Nevertheless, because PISA has been shown to be representative of the general population of 15 year olds, our study remains externally valid.

[14] See Table 3.

4.3.1 Summary Statistics

Table 3: **Descriptive Statistics showing Mean and SD for PISA 2012-2018**

	2012				2015				2018			
	Mathematics		Lang. Arts		Mathematics		Lang. Arts		Mathematics		Lang. Arts	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Standardised test score	-0.00	1.00	0.00	1.00	-0.00	1.00	0.00	1.00	-0.00	1.00	-0.00	1.00
Length in hours of avg. class	0.88	0.24	0.87	0.24	0.85	0.24	0.85	0.24	0.89	0.29	0.89	0.29
Weekly number of class periods	4.35	1.44	4.22	1.42	4.45	1.63	4.22	1.55	4.41	1.60	4.18	1.49
Weekly hours of instruction	3.77	1.52	3.64	1.48	3.73	1.70	3.55	1.65	3.84	1.85	3.65	1.76
Standardised Student-Related class quality	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	-0.00	1.00	-0.00	1.00
Standardised Teacher-Related class quality	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	-0.00	1.00	-0.00	1.00
Standardised Overall class quality	-0.00	1.00	-0.00	1.00	-0.00	1.00	-0.00	1.00	-0.00	1.00	-0.00	1.00

Notes. This table summarizes descriptive statistics for student performance, instruction time, and classroom quality indexes across the 2012, 2015, and 2018 PISA cycles. Values are rounded to 2 decimal places. "Lang. Arts" denotes Language Arts. The sample consists of 49 countries: Australia, Belgium, Brazil, Bulgaria, Canada, Chile, Colombia, Costa Rica, Croatia, Czechia, Denmark, Estonia, Finland, France, Germany, Great Britain, Greece, Hong Kong, Hungary, Ireland, Israel, Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Macao, Mexico, Montenegro, Netherlands, New Zealand, Norway, Peru, Poland, Portugal, Qatar, Russia, Singapore, Slovakia, Slovenia, Spain, Sweden, Switzerland, Thailand, Turkey, United Arab Emirates, Uruguay, and United States of America. **Data source:** The OECD and PISA

In Table 3, we report the mean and standard deviation for multiple factors that are relevant to our analysis. We report these factors for mathematics and language arts, as these are the two subjects that our main analysis contains, as explained in Section 4.2.1. We report all values for all three PISA cycles that we are assessing, 2012, 2015 and 2018. The responses for test score, student-related, teacher-related and overall classroom quality have been standardised to a mean of 0 and a standard deviation of 1. This ensures comparability across all years and subjects for the regression analysis. We detail the final sample size for the student-related classroom quality analysis in Section 4.2.3. Below we detail the sample size for the teacher-related classroom quality analysis and the overall classroom quality analysis.

For teacher-related classroom quality:

- **2012**: The main sample for this year comprises 257,484 observations from 125,176 9th and 10th-grade students across 13,324 schools.
- **2015**: The main sample for this year comprises 515,888 observations from 299,706 9th and 10th-grade students across 10,848 schools.
- **2018**: The main sample for this year comprises 385,346 observations from 210,862 9th and 10th-grade students across 12,386 schools.

For overall classroom quality:

- **2012**: The main sample for this year comprises 255,096 observations from 123,910 9th and 10th-grade students across 13,190 schools.

- **2015**: The main sample for this year comprises 509,470 observations from 296,550 9th and 10th grade students in 10,706 schools
- **2018**: The main sample used for this year comprises 382,658 observations from 209,684 9th and 10th grade students in 12,290 school

Table 3 shows an overall stable allocation of educational resources across the three PISA cycles, where mathematics consistently receives slightly higher weekly instruction time and class periods compared to language arts.

5 Methodology

In this section, we outline our strategy for measuring the impact of class time and environment on academic results. We rely on the cleaned datasets from PISA 2012, 2015, and 2018 described in Section 4 to serve as the foundation for this empirical analysis.

5.1 Identification Strategy

Estimating the causal effect of instructional time and classroom quality on student performance is difficult because of confounding factors arising from non-random sorting. Rivkin and Schiman (2015) note that unobserved factors correlated with both instructional time and classroom quality can bias the results. As a result, simple correlations between time and test scores are likely to be misleading. In particular, three main sources of bias must be addressed: non-random sorting and tracking, endogenous family responses, and the opportunity cost of time.

Firstly, results may be biased due to the non-random allocation of instruction hours and academic tracking. High-achieving students are more likely to attend schools that offer a 'higher education track', whereas comparatively low-achieving students may opt for schools that provide a 'vocational track'. Schools that offer a more academically-oriented education are likely to have higher instructional hours in math, science and/or language arts. This means that results could be upward biased, attributing the higher scores of these students to extra hours rather than their underlying ability. Furthermore, reverse causality may contribute to this upward bias through positive selection mechanisms. High-achieving students are often sorted into academic tracks or advanced courses that mandate more instruction time than vocational or standard tracks (Rivkin and Schiman, 2015). Consequently, the observed correlation may reflect the fact that higher prior achievement leads to increased class time, rather than the additional time increasing the level of student achievement.

Secondly, our analysis must consider the potential endogenous family response. For instance, if parents feel that students are receiving too little instruction time in school,

they may try to 'substitute' for this lack of instruction time by hiring a private tutor and ensuring their child receives a sufficient amount of time-on-task. This renders it difficult to identify the effect of instructional hours on student performance, as students with low in-school instruction time may still perform well if they have out-of-school instruction time to compensate. The direction of bias is ambiguous in this case.

Lastly, a key challenge in identifying the effect of instructional time lies in understanding its opportunity costs. For instance, when a school adds one additional hour to mathematics class time, it is plausible that they are removing instruction time elsewhere as there is only a limited amount of in-school instruction time per day. It may be other academic subjects, non-academic subjects (e.g., physical education), or even an extension of the school day, which reduces non-school time and possibly, out-of-school instruction time. Therefore, there is an opportunity cost no matter what.

Figure 1: **Correlations Between Weekly Hours of Instruction in Mathematics and Other Subjects by Country for PISA 2012-2018**.

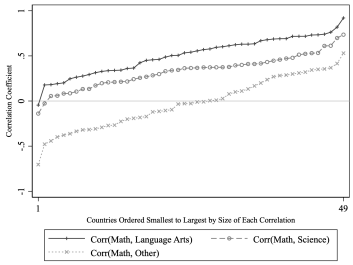


Figure 1.1: PISA 2012

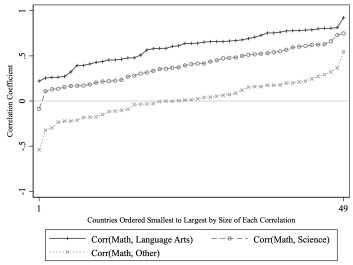


Figure 1.2: PISA 2015

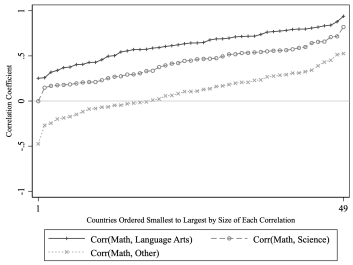


Figure 1.3: PISA 2018

Notes. This figure shows the correlation between weekly Mathematics instruction hours and time spent on other subjects for the 49 countries in our sample across PISA 2012, 2015, and 2018. "Other" subjects are calculated as the total weekly classes minus those for Mathematics, Language Arts, and Science. To ensure accuracy, we excluded invalid data, such as students reporting more than 50 classes per week. **Data source:** The OECD and PISA.

As depicted in Figure 1, the vast majority of countries show a positive correlation between time spent on mathematics and time spent on other subjects. This pattern suggests that additional time in mathematics is usually found in schools with longer school days or schools that devote more time to academic subjects in general, as opposed to non-academic subjects. Interestingly, from 2012-2018, it seems like there has been an upward shift in the correlations, particularly between mathematics and 'other' subjects. This could indicate that schools are increasingly adopting a model where there is more mathematics instruction time, paired by instruction time in other subjects - potentially achieved by longer school weeks in general.

Figure 1 highlights why a simple cross-school comparison will not help in investigating our research question. If we simply compare schools with more mathematics instruction time to less mathematics instruction time, we would likely be comparing systematically different types of schools.

In the next section, we discuss how we address these identification challenges by employing two different fixed effect specifications: the school-by-grade fixed effect and the school-by-subject fixed effect.

5.2 Implementation of the Identification Strategy

We adopt an empirical strategy following Rivkin and Schiman (2015), who demonstrated that exploiting the nested structure of PISA data effectively mitigates biases associated with non-random sorting of instruction time and unobserved effects across students and schools.

We implement panel data methods to exploit the fact that PISA tests multiple subjects for every student. The depth of the PISA dataset allows us to do this, which helps in controlling for unobserved student and school factors across subjects (Rivkin and Schiman, 2015). In addition to this, Lavy (2010) and Dee and Cohodes (2008) also show support for this method by showing in their research that panel data methods account for unobserved student and school characteristics across schools. This means that unobserved characteristics, such as teaching quality or innate mathematical ability, are effectively accounted for by adopting this empirical strategy.

One of the main hypotheses we will be testing is whether extending class time in a class with high classroom quality (student-related and teacher-related) will have higher benefits on student achievement.[15]

5.2.1 Baseline Empirical Model

We begin with a simple linear model to estimate the baseline return to instruction time (λ):

$$A_{kgsc} = \lambda M_{kgsc} + \alpha_s + \beta_g + \gamma_k + \delta_c + \theta_{kgc} + \omega_{gs} + \tau_{ks} + \varepsilon_{kgsc} \tag{1}$$

The variables of interest in our baseline specification are A_{kgsc} and λM_{kgsc}. A_{kgsc} represents the test scores in a specific subject k, in a specific grade g, in a specific school s, for a specific country c. A_{kgsc} is the dependent variable in all the regressions we run. M_{kgsc} represents the minutes of instruction time per week for a specific subject, in a specific grade, in a specific school, for a specific country. When we multiply M_{kgsc} with λ (λM_{kgsc}), we get our estimated coefficient that tells us how much an additional minute of class time per week affects student test scores.

Following our variables of interest, we add a series of fixed effects. α_s is the school fixed effect. It captures the differences between all schools in our sample, regardless of the subject. Likewise, β_g is the grade fixed effect. It essentially accounts for the fact that 10th graders are likely to score better than 9th graders, simply because of the added

[15] The list of all hypotheses being tested in this study can be found in Section 2.5.

advantage of age, maturity and cumulative learning. Similarly, γ_k is the subject fixed effect. This accounts for varying levels of difficulty between the tests of different subjects. For instance, the mathematics test might be inherently harder than the language arts test, which is accounted for by this fixed effect. The country fixed effect, which captures national trends due to culture and economy, is represented by δ_c. This might include cultural norms regarding the value of education, the structure of the national curriculum, or the general level of economic development of a country.

Moving forward, we include a set of interaction fixed effects to ensure that more specific biases are removed from our coefficient estimates as well. We start with θ_{kgc} which is the country-by-grade-by-subject fixed effect. It helps in accounting for national policies that affect factors like difficulty of a particular subject for a particular grade. It removes the bias of national curriculum and its varying standards across countries. We then move to the two most important fixed effects of our empirical model, which are followed by ε_{kgsc} which is the residual error term.

Our identification relies on two key specifications, which are included one at a time across the different regression models, to isolate the effect of instruction time (λ) by absorbing different layers of school and student characteristics:

1. School-by-Grade Fixed Effect - ω_{gs}

   This is the primary model that identifies the effect of instruction time based on differences in minutes and class periods between subjects within the same grade and school. The ω_{gs} fixed effect removes any factor common to students in a specific grade and school. This includes overall school quality, the average ability of the grade cohort, family background, and national funding structures. Therefore, the coefficient estimate that is produced by this specification is likely to be upward biased as it relies purely on within-school, within-grade, across-subject variation. In other words, it gives us a 'ceiling' estimate for our coefficients. This is because the model fails to take into account the fact that the difference in instruction time between subjects is often positively related to the difference in average ability or quality of instruction in those subjects.

2. School-by-Subject Fixed Effect - τ_{ks}

   The school-by-subject specification helps us in addressing the upward bias introduced by the school-by-grade model. Here we replace ω_{gs} with τ_{ks}. This specification takes into account differences in instruction time between grades within the same school and subject. For example, comparing 9th-grade mathematics minutes to 10th-grade mathematics minutes within a school. The τ_{ks} fixed effect accounts for those stable, subject-specific differences among schools, such as subject-specific teacher quality or national curricula. It is important to note that the strict exogeneity

assumption is violated as the model assumes that 9th-grade instructional time has no impact on 10th-grade achievement (Meghir and Rivkin, 2011). This means that this specification causes noisier results as they are likely to suffer from measurement error that leads to attenuation bias. Therefore, the school-by-subject fixed effect likely underestimates the correlation between instruction time and student achievement. This means that the coefficient estimates produced from this specification are 'floor' estimates.

Therefore, we interpret these results as an interval estimate, where the true impact of instruction time falls somewhere between the 'ceiling' produced by the school-by-grade model and the 'floor' produced by the school-by-subject model. Finally, we run all regressions separately for each year in our sample due to the data inconsistencies between the cycle-to-cycle datasets mentioned in Section 4.2.6. This approach also accounts for the varying measurement precision of student achievement, which fluctuates as the major domain rotates between mathematics (2012), science (2015), and language arts (2018). Having established the baseline empirical model which isolates linear effects, we move on to discuss dimensions of heterogeneity and how we capture those.

5.3 Dimensions of Heterogeneity

We examine two key dimensions of heterogeneity: the diminishing returns of additional instruction time on student achievement and the extent to which classroom quality influences the impact of additional instructional time on student achievement.

Firstly, we examine whether additional instruction time leads to diminishing returns on instruction time, and hence, smaller gains in learning. This may happen due to factors such as student fatigue or loss of concentration. This underscores that, even with comparable total weekly instructional time across schools, variations in class frequency and length can influence learning outcomes differently.

By using information on both the number of classes per week and the average class duration, we are able to assess whether additional instructional time exhibits diminishing returns. In practice, class lengths for our years vary little across subjects, consistent with Rivkin and Schiman (2015). This means our analysis is focused on the number of classes and total instruction minutes per week. We employ quadratic and higher-order terms to formally investigate the presence and magnitude of diminishing returns.

Moving forward, to examine how the learning environment influences returns to instruction time, we utilise the interaction model proposed by Rivkin and Schiman (2015). We construct a classroom quality index from administrator responses regarding student behaviour and teacher behaviour and interact both these measures separately with our instruction time variables. Table 4 and 5 list the variable factor loadings for our student and teacher-related classroom quality indexes respectively. The factor analysis conducted for

both student-related and teacher-related classroom quality reveals a strong correlation between various student behaviours (e.g., *student truancy* and *student skipping*) and teacher behaviours (e.g., *teacher absenteeism* and *teachers not being well-prepared*) meaning we cannot isolate the effect of individual factors like disruption or absenteeism. Consequently, we combine them into two separate indexes. We also re-scale the responses so that higher values indicate a higher-quality and less disruptive environment for both indexes.

Table 4: **Factor Analysis of Questions on Student Behaviour (2012-2018)**

	Student-Related Classroom quality					
	2012		2015		2018	
	Factor loadings	Scoring coeff.	Factor loadings	Scoring coeff.	Factor loadings	Scoring coeff.
Lack of student truancy	0.7959	0.2736	0.7793	0.2425	0.8114	0.2350
Lack of student skipping	0.7942	0.2820	0.8085	0.2974	0.8196	0.2614
Students respect teachers	0.7079	0.1651	0.7427	0.1903	0.7921	0.1771
Lack of student drug/alcohol use	0.7289	0.1923	0.7409	0.2013	0.8079	0.2158
Lack of student bullying	0.7250	0.2464	0.7088	0.2183	0.7835	0.2388

Notes. Factor analysis was conducted at the school level. The resulting index is standardised (Mean=0, SD=1), with higher values representing higher student-related quality classroom environments. Scoring coeff. refers to Scoring Coefficient. **Data source:** The OECD and PISA

Table 5: **Factor Analysis of Questions on Teacher Behaviour (2012-2018)**

	Teacher-Related Classroom quality					
	2012		2015		2018	
	Factor loadings	Scoring coeff.	Factor loadings	Scoring coeff.	Factor loadings	Scoring coeff.
Teachers not too strict with students	0.6553	0.1547	0.6488	0.1549	0.7084	0.1619
Teachers meeting students needs	0.7637	0.2362	0.7497	0.2282	0.7961	0.2391
Lack of Teacher absenteeism	0.7071	0.1855	0.7005	0.1856	0.7650	0.2041
Lack of Staff resisting change	0.7889	0.2772	0.7796	0.2717	0.8097	0.2644
Teachers well-prepared	0.7890	0.2744	0.7940	0.2894	0.7964	0.2449

Notes. Factor analysis was conducted at the school level. The resulting index is standardised (Mean=0, SD=1), with higher values representing higher teacher-related quality classroom environments. Scoring coeff. refers to Scoring Coefficient. **Data source:** The OECD and PISA

The factor analysis results in Tables 4 and 5 demonstrate that every variable shows a strong positive correlation with classroom quality (either student or teacher) across all PISA cycles. For example, *lack of student skipping* shows a factor loading[16] of around 0.8 across all years, and *teachers well-prepared* falls around 0.79 across all cycles. Therefore, we keep all of variables to create a robust and comprehensive index of student and teacher-related classroom quality. Furthermore, the consistently high factor loadings indicate that these variables are highly correlated to one another. This means that we cannot separate the variables within the index due to multicollinearity. This is because behaviours such as *lack of student truancy* and *students respect teachers* are likely to move together, meaning putting them into a regression model individually would result in unstable or insignificant

[16]The factor loadings indicates how strongly a specific question relates to classroom quality, wheras the scoring coefficient is a weight that is automatically calculated in our factor analysis and determines how much the specific question should count towards building the classroom quality index.

coefficients as they would compete for statistical significance. By combining these highly correlated variables into a single index, we mitigate this multicollinearity.

From Table 4 and 5 we can see that there is a high degree of stability in the factor loadings across the PISA cycles. This is interesting as it means that despite the numerous changes that occurred during this time period, such as rise of ICT in the classroom and smartphones among other factors discussed in the introduction, what influences the learning environment in a classroom has remained relatively consistent. This suggests that the definition of classroom quality remains the same over these PISA cycles. Although the variable with the highest factor loading changed from year-to-year, the overall proportions and magnitudes of the factor loadings remain largely similar across all cycles. Notably, this aligns with the relatively consistent pattern that can be observed in Tables 1 and 2.

5.4 Discussion of the Identification Strategy

Contemporaneous specifications, like ours where we measure how current class time impacts current scores, are biased if current inputs are correlated with the unobserved ability of the student or if schooling and family history is missing (Todd and Wolpin, 2003). For example, if smarter students (those with high unobserved ability) are placed in longer classes by their schools, or if wealthy families (unobserved family history) hire tutors, a simple regression will think the class time caused the high score when really it was the student's ability or the family response to instruction time. Todd and Wolpin (2003) warn that this leads to biased results. Since PISA data is cross-sectional, we cannot use a value-added approach which would allow us to compare a student's score this year to their score last year to account for cumulative learning.

Instead, we adopt the fixed effects strategy to absorb the unobserved endowment identified in Todd and Wolpin (2003) as a primary source of bias. The school-by-subject fixed effect model compares math and language arts for the same student cohort. By comparing how the same group of students performs in mathematics versus language arts, our regression is able to cancel out their unobserved ability. This means that if a student cohort is naturally smarter, they would perform better in both subjects. If they score higher in mathematics than language arts, and they also have more mathematics classes than language arts classes, we can be more confident it is the added instruction time causing the difference and not their innate ability.

Furthermore, we use the interaction term of instruction time and classroom quality to address the following problem: if you measure inputs wrong, the results will be biased. Therefore, by interacting instruction time with classroom quality we measure time-on-task rather than just scheduled time which reduces the measurement error that Todd and Wolpin (2003) warn leads to attenuation bias.

6 Results

In this section, we present our estimates of the effects of instructional time and classroom quality on PISA outcomes. We employ the fixed-effects models outlined in our methodology to analyse these relationships. We measure instructional time in two ways: the number of classes per week and total weekly hours.

We begin by examining the within-school variation in class time to validate our approach. We then report the average effect of instructional time on student achievement. Finally, we investigate whether the benefits of additional instructional time diminish as hours increase and how these benefits are influenced by the quality of the classroom environment.

6.1 Instruction Time Differences Between Subjects

In Table 6, we describe the distribution of weekly language arts instruction time conditional on mathematics instruction time for both total weekly minutes and the number of classes for 2012, 2015 and 2018[17]. The aim of this conditional probability table is to show that instruction time variation exists within schools between mathematics and language arts. We see that although the diagonal elements have the highest frequencies, meaning that many students tend to have very similar amounts of class time in both subjects, a substantial share of schools still report instruction time disparities between subjects.

[17] See Table A3 in Appendix on Page 46-47.

Table 6: **Distributions of Weekly Language Arts Instructional Minutes and Classes Conditional on Mathematics Instruction Time for (2012- 2018)**

Table 6.1: PISA 2012

Minutes per week	Mathematics							
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+	Observations
0–99	0.434	0.050	0.024	0.010	0.024	0.011	0.016	5,116
100–179	0.383	0.562	0.158	0.231	0.120	0.095	0.069	26,690
180–199	0.084	0.132	0.498	0.011	0.181	0.100	0.037	19,881
200–219	0.023	0.110	0.007	0.604	0.008	0.155	0.056	17,473
220–239	0.018	0.075	0.138	0.006	0.520	0.069	0.031	15,462
240–279	0.017	0.041	0.117	0.085	0.089	0.450	0.173	22,417
280+	0.040	0.030	0.057	0.053	0.057	0.120	0.617	21,469
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Observations	5,480	20,860	20,091	15,212	15,993	27,407	23,465	128,508

Table 6.2: PISA 2015

Minutes per week	Mathematics							
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+	Observations
0–99	0.472	0.063	0.031	0.022	0.046	0.024	0.032	16,713
100–179	0.314	0.594	0.165	0.265	0.108	0.110	0.072	56,651
180–199	0.091	0.149	0.556	0.002	0.212	0.101	0.031	43,926
200–219	0.015	0.090	0.002	0.577	0.000	0.172	0.049	29,944
220–239	0.021	0.057	0.149	0.000	0.486	0.065	0.032	29,912
240–279	0.017	0.032	0.073	0.093	0.081	0.434	0.143	38,978
280+	0.069	0.015	0.024	0.041	0.066	0.094	0.641	41,308
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Observations	16,838	42,786	39,783	25,088	33,549	52,002	47,386	257,432

Table 6.3: PISA 2018

Minutes per week	Mathematics							
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+	Observations
0–99	0.487	0.061	0.031	0.021	0.032	0.016	0.027	11,021
100–179	0.289	0.592	0.144	0.248	0.115	0.103	0.071	39,647
180–199	0.084	0.139	0.560	0.002	0.155	0.100	0.037	29,507
200–219	0.015	0.097	0.002	0.620	0.000	0.169	0.052	25,067
220–239	0.023	0.056	0.158	0.000	0.583	0.066	0.053	27,561
240–279	0.023	0.035	0.077	0.073	0.062	0.455	0.128	27,818
280+	0.079	0.019	0.027	0.037	0.053	0.091	0.632	31,671
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Observations	11,159	29,087	26,747	22,588	29,356	35,568	37,787	192,292

Notes. Here, we present the joint distribution of students by separate groupings of weekly mathematics and language arts instruction time measured in minutes per week. Please refer to Table A3 on Page 46-47 in the Appendix for the full table, which includes the distribution of class periods. **Data source:** The OECD and PISA

In Table 6, the 200-219 minutes column (mathematics) matched with the 200-219 minutes row (language arts) provides probabilities for students to have the same amount of class time for both subjects. The given probabilities are 0.604 for 2012, 0.577 for 2015 and 0.620 for 2018. This can be read as: of the students who have 200-219 minutes of mathematics per week, 60.4% of them also have the exact same amount of language arts in 2012. With this pattern we see that the diagonal from the top left to bottom right has the highest frequencies, meaning schools generally allocate similar amounts of instruction time to mathematics and language arts.

However, the numbers on the right and left of the diagonal show greater disparities between instruction time. This variance in the misalignment between instruction time between the two subjects is what we exploit to see whether the subject that students spend more time in actually gets better scores. Across all three PISA cycles, the data reveals a distinct asymmetry in how instructional time is allocated when observing the upper and

lower tails of the distribution. For example, we see that substitution between mathematics and language arts is common at the lowest levels of mathematics instruction. For example, 43.4% of the students with 0-99 minutes for mathematics also receive 0-99 minutes of language arts time. However, 38.3% receive 100-179 minutes of language arts time. This suggests that for students with very low mathematics exposure, schools may be streaming students into language arts rather than mathematics education. On the upper tail we do not see the reverse trade-off. This means that students receiving 280+ minutes of mathematics instruction are unlikely to receive low language arts instruction. This indicates that high instructional time is complementary, schools that invest heavily in mathematics tend to also invest heavily in language arts rather than trading off one subject for the other. This pattern is consistent across 2012-2018, reinforcing that the resource constraint trade-off is visible mostly at the lower end of the distribution.

We also observe that differences in instruction time between categories tend to be 20 minutes per week or more. This pattern is consistent with the nature of school schedules where the average duration of a single class period typically exceeds 40 minutes [18]. A similar pattern holds for classes per week, seen in Table A4[19], which is the primary source of within-school instruction time variation. Students who attend three mathematics classes per week are far more likely to attend four or more language arts classes than fewer than three. Although, Table 6 shows the distribution of class time, it does not know where this class time originates. For example, a higher amount of instruction time for one or both subjects could be due to remedial classes for low-achieving students to help them catch up or due to streaming where high-achieving students are given more class in the subject where they excel. In both of these cases, knowing this information changes our understanding of the results.

Table 6 documents the existence of adequate within-school instruction time variation to identify effects. In Table A4 we describe patterns of test score differences to examine whether the raw test score data are consistent with the belief that longer classes raise achievement. Table A4 reports differences in average standardised test scores[20] by the distribution of mathematics and language arts instruction time based on both minutes and classes per week. Average test scores are standardised to have a mean of 0 and a standard deviation of 1 for each grade and subject. This table has the same structure as Table 3 but replaces the conditional distributions with the average standardised score differences.

Consistent with the hypothesis that instruction time leads to achievement, we find that entries above the diagonal (where instruction time for mathematics exceeds that for language arts) tend to be more positive than entries along the diagonal, where instruction time is balanced. In turn, these diagonal entries tend to be more positive than those below the diagonal, where instruction time for language arts exceeds that for mathematics. This

[18] We find that between 2012-2018 the average length of a class is between 51-53 minutes, as seen in Table 3

[19] See Page 47-48 in Appendix.

[20] Calculated by subtracting standardised language arts scores from standardised mathematics scores in the PISA assessments.

pattern implies that the more the relative time spent on mathematics increases, the relative performance in mathematics improves. Entries along the diagonal, where time allocation is roughly equal, tend to hover closer to 0 compared to the extremes, serving as a neutral baseline that further validates the marginal impact of shifting time allocation from one subject to the other. In the data for 2015 we see some noise as there are some large positive values below the diagonal. However, the overall directional trend across the three PISA cycles confirms that shifting time allocation toward mathematics is associated with a relative increase in mathematics achievement.

6.2 Instruction Time Effects

This section examines how instruction time affects student achievement. We first present estimates from basic linear models, followed by models that allow for a more flexible relationship between time and test scores. All tables include coefficients with school-by-grade and school-by-subject fixed effects, along with standard errors clustered at the school level. In every regression, the dependent variable is the average PISA standardised score for each school-grade-subject cell. The main sample consists of 256,316 school-grade-student-subject observations in 2012, in 2015 there are 513,778 observations, and 382,658 observations in 2018. When using the school-by-subject specification, approximately 51% of observations are dropped in 2012, approximately 42% in 2015, and approximately 45% in 2018.

Table 7: **Estimated Effects of Weekly Hours of Instruction and Class Periods on Achievement (2012–2018)**

	2012			2015			2018		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Panel (a): Weekly Hours of Instruction Time									
Coefficient	0.068***	0.027***	0.008	0.087***	0.007	0.013*	0.030***	0.018***	-0.004
	(0.008)	(0.005)	(0.008)	(0.010)	(0.006)	(0.008)	(0.009)	(0.006)	(0.007)
Panel (b): Weekly Number of Class Periods									
Coefficient	0.106***	0.030***	0.019**	0.139***	0.007	0.032***	0.129***	0.018***	0.040***
	(0.008)	(0.005)	(0.009)	(0.011)	(0.005)	(0.008)	(0.012)	(0.005)	(0.009)
School-by-grade FE	N	Y	N	N	Y	N	N	Y	N
School-by-subject FE	N	N	Y	N	N	Y	N	N	Y
Sample Size	256316	256316	124456	513778	513778	299312	382658	382658	209684
No. of Schools	13263	13263	5439	10801	10801	5369	12290	12290	5596

Notes. Robust standard errors clustered at school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable in all regressions is the school-by-grade-by-subject average standardised test score based on the PV1MATH and PV1READ variables. Estimates are insensitive to the choice of plausible value. All regressions also include a country-by-grade-by-subject fixed effect. The samples are limited to students who have both mathematics and language arts components in their PISA test booklets and valid class-time information for both subjects. Robust standard errors clustered by school are in parentheses. **Data source:** The OECD and PISA.

Starting with 2012, in the baseline specifications without school fixed effects that can be observed in column 1, we observe a strong, positive relationship between instruction quantity and test scores: a one-hour increase in weekly instruction is associated with a

0.068 SD rise in student test scores, compared to a 0.106 SD rise for each additional class period. The estimated effects of instruction time decline as we introduce more restrictive fixed effects as reported in column 2 and 3 for 2012 in Table 7. When we include school-by-grade fixed effects in column 2 to control for differences in school quality and student characteristics across subjects, the coefficient for weekly hours of instruction time falls by around 60%, while the coefficient for weekly number of class periods falls by around 72%. However, these estimates are likely upward biased as discussed in Section 5.2.1. This suggests that a large amount of the raw correlation between instructional hours and student achievement is driven by unobserved school and student factors.

The decline in the coefficient magnitude and statistical significance when moving from the school-by-grade fixed effect specification to the school-by-subject fixed effect model, in other words, from column 2 to column 3, is consistent with the presence of learning persistence between 9th-grade and 10th-grade. As explained in Section 5.2.1, instruction time in the current grade is often correlated with instruction time in the previous grade. The school-by-subject specification fails to take this into account by trying to isolate just one year's contribution by differencing grades. However, because learning persistence still exists and this cannot be measured by this model, the strict exogeneity assumption is violated as the model assumes that 9th-grade instructional time has no impact on 10th-grade achievement. This means the results become more noisy in column 3 for 2012, and the estimated coefficients suffer from measurement error that leads to attenuation bias. They likely underestimate the correlation between instruction time and student achievement. To ensure the noisy results are not stemming from a lack of variation in the data, in Table A5[21] in the Appendix we observe that the school-by-subject specification for 2012 in column 4 and the school-by-grade specification for 2012 in column 3 for both weekly hours of instruction and weekly number of class periods have roughly 10% of useful variation left to exploit, further supporting the learning persistence theory outlined above. In addition, the estimated correlation between 9th-and 10th-grade differences for the year 2012 is 0.38 which further substantiates the idea that knowledge carries over from one year to the next. Therefore, the school-by-grade specification's coefficient overestimates the effect of additional test scores on student achievement, while the school-by-subject specification underestimates it. These findings mirror Rivkin and Schiman (2015). Using 2009 PISA data, they demonstrate that the return on an extra hour of class is small (0.023-0.031 SD) when student and school quality are properly accounted for.

At this stage, our estimates depart from the patterns observed in 2012. In the 2015 sample, estimates from the baseline specification are particularly large in magnitude and statistically significant at the 1% level. However, in Table 7 column 2 for 2015, we can see that both coefficients (0.007 for weekly hours and 0.007 for class periods) are not statistically significant and are small in magnitude. However, the 2015 results show a notable

[21] See Page 49 in Appendix.

increase in magnitude within the school-by-subject specification. Under this specification, a one hour increase in instruction time predicts a 0.013 SD rise in scores, and an additional class period predicts a 0.032 SD rise, with significance at the 10% and 1% levels, respectively.

Why might this be happening? The estimated correlation between 9th-and 10th-grade differences for the 2015 data is 0.48 which means that learning persistence exists between 9th-graders and 10th-graders in our dataset to an even stronger degree. However, 2015 was the first time the test was also offered in a computer-based format as mentioned in Section 4.2.6. This means that a student's score was no longer just a function of their subject-skill in the school-by-grade specification, but also of their ICT familiarity. Therefore, a potential explanation for this divergence in pattern in 2015 is that a new source of error is introduced in the form of the student's familiarity and proficiency in navigating a computer- interface test. This is especially true for countries, such as Japan, Korea and Hong Kong, who historically have not had much digital integration into the education system (Komatsu and Rappleye, 2017). This is another reason that could cause measurement error in the model. Likewise, it could also be specific cohort characteristics for the year 2015 that we cannot observe. The school-by-subject model aggregates data to the grade-level, which means differences in a student's ability to use a computer mouse for example, were likely averaged out to a certain degree which could be why we see coefficient estimates of a larger magnitude that are also statistically significant. The statistical significance of the coefficient estimates in column 3 for 2015 can also be attributed to the large sample size in that year. While 2012 includes 256,316 observations, the 2015 sample comprises 513,778 observations, more than double that of 2012. Statistically speaking, larger sample sizes reduce standard errors, which makes it easier to detect very small effects as statistically significant - this may be what is happening here as well. Likewise, as mentioned in Section 4.2.6, science was the main domain for the year 2015. This means that mathematics and language arts sections were shorter and given less attention. Because science was the main testing domain in 2015, most of the testing time[22] was allocated to science items. As a result, mathematics and language arts became minor domains that received far fewer test items than in 2012. This creates two issues for our analysis. Firstly, having fewer items reduces the reliability of the test scores for mathematics and language arts, since there is less information available to estimate student ability. Secondly, because science was prioritised, mathematics and language arts items were often placed in the second half of the two-hour session, which raises concerns about fatigue affecting performance as outlined by Nagy et al. (2018) in a Germany-specific study. For these reasons, the estimates for 2015 should be interpreted with caution, as they are based on fewer items and may be affected by test-taking fatigue.

[22] 60 minutes per student.

As mentioned in Section 4.2.6, 2018 was the first time a multi-stage adaptive test of reading was introduced. This seems to have partially restored the trend that could be observed in our 2012 results for Table 7. The baseline specification in column 1 holds the most statistically significant coefficient estimates that are of the largest magnitude for 2018. The coefficient for weekly hours of instruction time on student achievement falls by about 40% from the baseline model to the school-by-grade specification in column 2 for 2018. Likewise, the decrease is about 86% for the coefficient for weekly number of class periods on student achievement. The increase in precision seems to have amplified in the school-by-subject specification for 2018 in column 3, but only for Panel b. Perhaps this is because due to multi-stage adaptive testing, 10th-graders were given harder questions than 9th-graders which accounts for learning persistence to a certain degree and helps in eliminating some of the attenuation bias that was previously present in this specification. The fact that the weekly hours coefficient in Panel a drops to a negative, insignificant estimate may imply that the accuracy of this specific variable has deteriorated. Overall, this makes the estimates for the school-by-subject specification noisier for 2018, which can be observed visually in Figure 2.3.

Supplementary analysis that includes science scores in Table A6[23] shows that the effects of instructional time are stable over time. When the sample accounts for the major domain in 2015 which is science, the coefficient for weekly instruction hours is 0.028 SD in column 3, almost identical to the estimates for 2012 (0.029 SD) and 2018 (0.029 SD). This supports our explanation that the earlier variation was caused by the exclusion of the primary assessment domain.

Figure 2 (see Page 27) visualizes the relationship between instruction time and achievement for each year. We plot the residual variation in test scores against the residual variation in instruction time. These residuals represent the specific fluctuations in time and achievement that remain after we control for the school-by-grade (Panel a) and school-by-subject (Panel b) fixed effects. As can be seen, the relationship is stronger in Panel a for both 2012 and 2018 as expected, but this pattern does not hold true for 2015. The possible reasons for this are discussed within the same section.

We now investigate the possibility of diminishing returns to additional hours. Table 8 (see Page 28) reports results from the three specifications with weekly hours reported as a quadratic. The results in both fixed effect specifications strongly support the hypothesis of diminishing returns across all three PISA cycles. Specifically, while the coefficient on the linear term (Weekly Hours) is consistently positive and significant, the coefficient on the squared term is consistently negative and significant. This pattern holds true for 2015 and 2018, confirming that while additional instruction time improves achievement, the marginal benefit of each additional hour decreases as the total time increases.

[23] See Page 49 in Appendix.

Figure 2: **Instruction Time Residuals Plotted Against standardised Achievement Residuals**

Figure 2.1: PISA 2012

Panel (a): School-by-Grade Fixed Effects



Panel (b): School-by-Subject Fixed Effects



Figure 2.2: PISA 2015

Panel (a): School-by-Grade Fixed Effects



Panel (b): School-by-Subject Fixed Effects

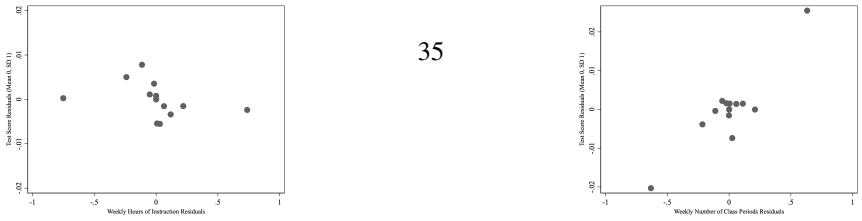


Figure 2.3: PISA 2018

Panel (a): School-by-Grade Fixed Effects



Panel (b): School-by-Subject Fixed Effects



Notes. Following the non-parametric method of Chetty et al. (2011), these binned scatterplots display achievement residuals against instruction time residuals. We first regress both variables on country-by-grade-by-subject fixed effects and either school-by-grade (Panel a) or school-by-subject (Panel b) fixed effects. We then bin the residuals and plot the means of these bins to visualize the relationship. **Data source:** The OECD and PISA.

Table A7[24] helps in indicating that parental compensation is not causing our results. Once we account for out-of-school lessons, the instruction-time coefficients stay stable or grow, suggesting that the advantages of classroom learning operate independently of private tutoring.

The purpose of Table A8[25] is to test whether the relationship between instructional time and achievement remains consistent across different levels of weekly minutes and class counts. Overall, both fixed-effect specifications show the expected positive pattern, but the top category of weekly minutes produces unexpected and negative estimates in some years. These inconsistencies suggest that extremely high reported minutes likely capture measurement error or remedial track students rather than true instructional gains. At the same time, the positive effects for higher numbers of weekly classes in 2018 indicate that the broader relationship between class frequency and achievement still holds.

Table 8: **Estimated Effects of Weekly Instructional Hours on Achievement from Quadratic Specifications for PISA 2012-2018**

	2012			2015			2018		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Weekly Hours of Instruction	0.3966***	0.1012***	0.0562***	0.5112***	0.0282*	0.0782***	0.2822***	0.0485***	0.0472***
	(0.0263)	(0.0150)	(0.0197)	(0.0312)	(0.0168)	(0.0219)	(0.0258)	(0.0120)	(0.0180)
Squared Weekly Hours of Instruction	-0.0356***	-0.0084***	-0.0048***	-0.0486***	-0.0027	-0.0069***	-0.0265***	-0.0035***	-0.0048***
	(0.0027)	(0.0015)	(0.0018)	(0.0035)	(0.0019)	(0.0022)	(0.0025)	(0.0012)	(0.0016)
School-by-grade FE	N	Y	N	N	Y	N	N	Y	N
School-by-subject FE	N	N	Y	N	N	Y	N	N	Y
Sample Size	256316	256316	124456	513778	513778	299312	384302	384302	210596
No. of Schools	13263	13263	5439	10801	10801	5369	12349	12349	5624

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

6.3 Heterogeneity by Classroom Environment

The idea that the return to additional time is dependent on classroom environment aligns with the Disruption Model shown in Lazear (2001). In this section, we examine the potential differences in the effect of instructional time based on student and teacher behaviours.

The data about classroom quality comes from the school questionnaire, where we receive one answer per school. Thus, the index does not vary within schools and the direct effect of classroom quality on student achievement cannot be identified, meaning we can only compare results across schools. In order to address this limitation, we measure if classroom quality changes the effectiveness of instruction time through the interaction term of classroom quality and instruction time. In this way, we can measure the heterogeneity in the returns to instruction time across the classroom quality index. We re-scale

[24] See Page 50 in Appendix.
[25] See Page 51 in Appendix.

the factors within the classroom quality index to have a mean of 0 and standard deviation of 1 across all countries and schools in our sample.

Table 9 measures whether the impact of weekly instruction time on student achievement changes depending on the quality of the student-related classroom environment. It shows if adding extra teaching time leads to higher test score gains in high-quality, with the interaction effect, compared to lower-quality classrooms, without the interaction effect. The table shows that the return to instruction time is heterogeneous with respect to the student-related classroom environment. The interaction coefficients for 2012 (0.0135 SD) and 2018 (0.00118 SD) are positive and statistically significant at the 1% level for the school-by-grade fixed effect, although modest in effect size. This indicates that adding an extra hour of instruction is more effective in classroom with better student behaviour (higher quality) then in those with poor behaviour. The estimate suggests that a 1 SD improvement in the student-related classroom quality index raises the return to an additional hour of instruction by 0.0135 SD in 2012. In 2015, the interaction term for school-by-grade is close to 0 (0.0012 SD) and in school-by-subject negative (-0.0027 SD) and not statistically significant at the 10% level for either estimate. This unusual set of results for 2015 is consistent with what we find in Table 7, and Figure 2.2 as well. There are a few different factors that could explain why we see these results for 2015 which are outlined in Section 6.2.

Table 9: **Estimated Effects of Weekly Instructional Hours and Class Periods on Achievement, by Student-Related Classroom Quality PISA 2012-2018**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours						
Weekly Hours of Instruction	0.0264***	0.0090	0.0065	0.0129*	0.0174***	-0.0033
	(0.0054)	(0.0081)	(0.0063)	(0.0077)	(0.0060)	(0.0075)
Weekly Hours × Student-related Classroom Quality	0.0135***	0.0064	0.0012	-0.0027	0.0118***	0.0074
	(0.0042)	(0.0063)	(0.0049)	(0.0069)	(0.0043)	(0.0068)
Panel (b): Weekly Classes						
Weekly Number of Class Periods	0.0285***	0.0201**	0.0067	0.0323***	0.0163***	0.0410***
	(0.0053)	(0.0090)	(0.0054)	(0.0083)	(0.0053)	(0.0090)
Weekly Classes × Student-related Classroom Quality	0.0126***	0.0102	-0.0005	0.0012	0.0082**	0.0072
	(0.0039)	(0.0072)	(0.0041)	(0.0076)	(0.0037)	(0.0084)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	257,016	124,858	513,778	299,312	384,584	210,772
No. of Schools	13,306	5,458	10,801	5,369	12,361	5,629

Notes. Robust standard errors clustered at school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

The coefficients for the school-by-subject fixed effect are generally smaller and less significant than school-by-grade fixed effect. The school-by-subject fixed effects absorb almost all variation. The results seen in this table are consistent to those seen by Rivkin and Schiman (2015). Due to this, the student-related classroom quality survey questions

are subjective and noisy, this leads to attenuation bias, which is likely why the effects disappear or shrink with this fixed effect.

Table A9[26] shows that the interaction-term coefficients become significant at the 1% level in the weekly hours and classes specifications with school-by-grade and school-by-subject fixed effects. Although, the magnitudes of the estimates remain quite small. We see that after including science, the sample size has increased and there is more data points to model the relationship with greater certainty. Thus, we see that by adding in science scores the noise in the data has been reduced. The fact that our results remain statistically significant with science being added indicates that the results are robust. Thus, Table A9 further supports our findings in Table 9.

The Disruption Model indicates there could be a non-linear relationship between the level of disruption in a classroom and time-on-task (Lazear, 2001). However, the categorical structure of the factors in our classroom quality index do not suggest any specific functional form. Thus, we estimate a specification that includes interactions between instruction time and quartiles of classroom environment in Table 8 as this allows us to be more flexible with the functional form of the data.

Table 10: **Estimated Effects of Instructional Time on Achievement, by Quartile of Student-Related Classroom Quality (PISA 2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Instructional Hours						
Instruction time	0.001	-0.014	0.001	0.015	0.008	-0.010
	(0.009)	(0.015)	(0.010)	(0.015)	(0.009)	(0.012)
Instruction time × 2nd quality student-related quartile	0.025**	0.056**	0.000	0.009	0.001	0.011
	(0.012)	(0.022)	(0.013)	(0.019)	(0.012)	(0.018)
Instruction time × 3rd student-related quality quartile	0.028**	0.007	0.023*	-0.013	0.005	-0.002
	(0.013)	(0.022)	(0.014)	(0.022)	(0.012)	(0.019)
Instruction time × 4th student-related quality quartile	0.045***	0.029	0.001	-0.006	0.031**	0.012
	(0.012)	(0.020)	(0.014)	(0.022)	(0.013)	(0.021)
Panel (b): Weekly Number of Class Periods						
Instruction time	0.007	0.001	0.003	0.022	0.012	0.037**
	(0.009)	(0.016)	(0.009)	(0.015)	(0.008)	(0.016)
Instruction time × 2nd student-related quality quartile	0.021**	0.043*	0.001	0.030	0.001	-0.007
	(0.011)	(0.024)	(0.011)	(0.020)	(0.010)	(0.022)
Instruction time × 3rd student-related quality quartile	0.022*	-0.005	0.017	0.001	-0.001	0.007
	(0.012)	(0.025)	(0.011)	(0.021)	(0.010)	(0.023)
Instruction time × 4th student-related quality quartile	0.042***	0.038*	-0.003	0.006	0.017	0.017
	(0.012)	(0.022)	(0.012)	(0.023)	(0.011)	(0.026)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	257,016	124,858	514,864	299,522	384,584	210,772
No. of Schools	13,306	5,458	10,831	5,374	12,361	5,629

Notes. Robust standard errors clustered at school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

[26] See Page 51 in Appendix.

In the 2012 school-by-grade specification, the results in Table 10 show a monotonically increasing return to additional instruction time as the quality of the classroom environment improves. The table presents a baseline effect size for instruction time, which has no interaction term, and then the marginal increases in returns to academic performance that we see for each specific quartile of classroom quality. For example, for the bottom quartile we see that the effect size is near zero (0.001). However, the the effect size rises as student-related classroom quality increases. For the second quartile the effect size is 0.025, meaning the total effect size is 0.001 + 0.025 = 0.026. Student in the top quartile of student-related classroom quality realise a benefit from additional hours that is substantially larger than those in the bottom quartile, with a total effect size of 0.001 + 0.045 = 0.046. Although the pattern for 2018 is not as drastic as 2012, we still see the same pattern hold.

The estimated benefit of additional class time is drastically higher for the top half of student-related classroom quality estimates. For 2012, the interaction coefficients for the top three quartiles all exceed 0.02 and are significant at the 1% level, while the the lowest quality quartile shows no significant return to added instruction time. Overall, these findings are consistent with the findings in Rivkin and Schiman (2015). The 2018 data suggests a threshold effect rather than a linear progression. The productivity of instruction time increases significantly only when the classroom environment reaches the highest quality quartile. This is because all values until then are not statistically significant, or even show negative estimates with school-by-subject fixed effects.

After adding science tests in Table A10[27], we find a slightly different result. When broadening the analysis to include science, the 2012 results reveal a distinct threshold effect, similar to 2018. The benefits of classroom quality in this robust specification are concentrated entirely in the top quartile. This suggests that a classroom must reach a high standard of student-related classroom environment before the complementarities between instruction time and academic achievement are prominent in the results. Including science scores reduces the standard errors and reveals that the results are statistically significant in the baseline model as well. This confirms that, on average, additional time does lead to gains in academic performance. As anticipated, the inclusion of science scores results in smaller interaction coefficients overall. This is consistent with presence of measurement error in the PISA test results leading to measurement error because mathematics and science skills are highly correlated.

As for the 2015 results, the interaction terms in the school-by-subject specifications are generally small and statistically insignificant across all years. This aligns with the expectation that controlling for subject-specific heterogeneity absorbs too much of the variation in the data and likely over corrects the model, particularly because the classroom quality index contain significant noise. As stated for Table 9, the difference in 2015 results

[27] See Page 52 in Appendix.

can likely be due to measurement error which would lead to noise in the results from the switch to computer-based testing which makes it harder to identify the effect of classroom environment.

6.4 Extension to the Study

In this section, we extend the methodology to account for alternative dimensions of the learning environment. Whereas Rivkin and Schiman (2015) isolated student-related factors, such as *student truancy*, we broaden the scope to examine teacher-related classroom quality, such as *teacher absenteeism*. Furthermore, to capture the relationship between student behaviour and teacher behaviour, we present results for a combined classroom quality index based on our individual student-related and teacher-related classroom quality indexes.

Table 11 (see Page 32) presents the estimates for teacher-related classroom quality, serving as the direct counterpart to the student-related analysis reported in Table 9. Upon first glance, it can be seen that none of the interaction terms for 2012 and 2015 (Weekly Hours x Teacher-related Classroom Quality) and (Weekly Classes x Teacher-related Classroom Quality) are statistically significant. Caution is required when interpreting these results because it is important to keep in mind that the PISA school questionnaire, while not directly answered by teachers, is answered by the school administrator or the principal in some cases. We address the potential reporting bias, as well as other possible explanations, in Section 7.5.

However, it is worth mentioning that in 2018, we see statistically significant interaction effects at the 10% level for both (Weekly Hours x Teacher-related Classroom Quality) and (Weekly Classes x Teacher-related Classroom Quality) in the school-by-grade specification in column 1. This aligns with our findings in Table 9, and suggests that in 2018, teacher-related classroom environment quality increased the effectiveness of instruction time modestly. According to Table 11, in the 2018 school-by-grade specification, the interaction coefficient of 0.0102 indicates that a 1 SD improvement in teacher-related classroom quality increases the effect of an additional hour of instruction by approximately 0.01 SD of student test scores. Similarly, for the number of class periods, the interaction coefficient suggests that a 1 SD improvement in teacher-related classroom quality raises the marginal benefit of an additional class period by approximately 0.009 SD of student test scores. As expected, the school-by-subject specification in column 2 for 2018 becomes smaller and statistically insignificant. Overall, the consistent patterns displayed by 2018 may be attributed to the fact that there is differing variation in the distribution of teacher-related responses for this year (see Table A2.2) as compared to the previous years.

Table 11: **Estimated Effects of Weekly Instructional Hours and Class Periods on Achievement, by Teacher-Related Classroom Quality (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Weekly Hours of Instruction	0.0269***	0.0099	0.0076	0.0120	0.0187***	-0.0037
	(0.0054)	(0.0082)	(0.0063)	(0.0078)	(0.0061)	(0.0076)
Weekly Hours × Teacher-related Classroom Quality	-0.0041	0.0072	0.0029	-0.0032	0.0102**	0.0010
	(0.0042)	(0.0068)	(0.0046)	(0.0069)	(0.0043)	(0.0065)
Panel (b): Weekly Number of Class Periods						
Weekly Number of Class Periods	0.0291***	0.0197**	0.0072	0.0306***	0.0176***	0.0404***
	(0.0053)	(0.0091)	(0.0054)	(0.0084)	(0.0054)	(0.0092)
Weekly Classes × Teacher-related Classroom Quality	-0.0018	0.0005	0.0025	0.0003	0.0087**	0.0002
	(0.0038)	(0.0076)	(0.0038)	(0.0076)	(0.0037)	(0.0077)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	257,484	125,176	515,888	299,706	385,346	210,862
No. of Schools	13,324	5,471	10,848	5,383	12,386	5,635

Notes. Robust standard errors clustered at school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

When the analysis is expanded to include science scores in Table A11[28], the interaction effect between class periods and classroom quality becomes robust and statistically significant across all three cycles. This consistency, contrasting with the limited findings in Table 11, is likely to be caused by the substantial increase in statistical power. The inclusion of science scores nearly doubles the sample size (e.g., increasing the 2012 sample from 257,484 to 481,098), thereby reducing standard errors and making it easier to detect interaction effects that were not visible in the smaller sample.

As mentioned in Section 6.3, there is the possibility that disruption affects learning non-linearly. We investigate this possibility in the same manner as Table 10 for teacher-related classroom quality (Lazear, 2001). Table 12 presents these results. In 2012, we see an unusual pattern where the baseline group has the highest return for the school-by-grade specification, and better teacher support actually lowers the marginal benefit of extra time relative to that baseline. Although this coefficient is statistically insignificant in Panel b and only statistically significant at the 10% level in Panel a. As expected, the statistical significance and magnitude of the coefficients decreases in the school-by-subject specifications. This unusual pattern continues in 2015, except the pattern can be observed in the school-by-subject specification rather than school-by-grade specification. This is perhaps due to lack of variation caused by the potential reporting bias as evidenced in Table 11 as well. In 2018, we get consistent results with Table 10 where in Panel a, the interaction with the top quality quartile indicates that students in the top quartile of teacher-related classroom quality experience a significantly larger marginal return to additional class periods compared to the bottom quartile, with an interaction coefficient

[28] See Page 52 in Appendix.

of 0.029 SD, resulting in a total effect size of 0.036 SD calculated as 0.007 + 0.029. This means that for students in the highest quartile of teacher-related classroom quality, an extra hour of weekly instruction is linked to a 0.036 SD increase in test scores for students while the effect for students in the lowest quartile is close to 0. Similarly, in Panel b, the interaction with the top quality quartile indicates that students in the top quartile of teacher-related classroom quality experience a significantly larger marginal return to additional class periods compared to the bottom quartile, with an interaction coefficient of 0.023 SD, resulting in a total effect size of 0.029 SD calculated as 0.006 + 0.023. Both these interaction coefficients are significant at the 10% level. This pattern disappears in the school-by-subject specification as expected, and the unusual pattern returns.

When the analysis is expanded to include science scores in Table A12[29], none of the interaction effects are statistically insignificant at the top quartile of teacher-related classroom quality. While this may indicate that science performance is not inherently linked to teacher behaviour and instructional support, one must be cautious when interpreting these results as the bias caused by the spillover effects between mathematics and science and reliability in the science domain may affect the precision of the estimates.

Table 12: **Estimated Effects of Instructional Time on Achievement, by Quartile of Teacher-Related Classroom Quality (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Instruction time	0.041***	-0.010	-0.001	0.027**	0.007	-0.001
	(0.009)	(0.014)	(0.011)	(0.013)	(0.010)	(0.011)
Instruction time × 2nd teacher-related Quality Quartile	-0.028**	0.038*	0.021	-0.026	0.006	-0.009
	(0.013)	(0.020)	(0.015)	(0.019)	(0.013)	(0.017)
Instruction time × 3rd teacher-related Quality Quartile	-0.008	0.022	0.014	-0.026	0.013	0.004
	(0.012)	(0.021)	(0.014)	(0.020)	(0.013)	(0.021)
Instruction time × 4th teacher-related Quality Quartile	-0.021*	0.023	0.003	-0.010	0.029**	-0.004
	(0.012)	(0.023)	(0.013)	(0.021)	(0.013)	(0.020)
Panel (b): Weekly Number of Class Periods						
Instruction time	0.039***	0.004	-0.000	0.038***	0.006	0.045***
	(0.009)	(0.015)	(0.009)	(0.013)	(0.009)	(0.015)
Instruction time × 2nd teacher-related Quality Quartile	-0.023*	0.037*	0.018	-0.014	0.010	-0.013
	(0.012)	(0.022)	(0.012)	(0.021)	(0.011)	(0.021)
Instruction time × 3rd teacher-related Quality Quartile	-0.002	0.017	0.013	-0.027	0.014	-0.017
	(0.011)	(0.023)	(0.012)	(0.020)	(0.011)	(0.024)
Instruction time × 4th teacher-related Quality Quartile	-0.016	0.013	0.003	0.013	0.023**	0.009
	(0.011)	(0.024)	(0.011)	(0.021)	(0.011)	(0.024)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	257,484	125,176	515,888	299,706	385,346	210,862
No. of schools	13,324	5,471	10,848	5,383	12,386	5,635

Notes. Robust standard errors clustered at school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

[29] See Page 53 in Appendix.

Finally, we present estimates using the combined classroom quality index, which aggregates both student-related and teacher-related factors to provide a broader measure of the learning environment. Table 13 reports the linear interaction effects of the combined classroom quality index and instruction hours. For 2012, Table 13 indicates only marginally significant interaction effects and mirrors the trend of Table 9. In column 1 of Panel b, the interaction coefficient is 0.0067 SD suggesting a weak interaction effect, as opposed to no interaction effect. This interaction coefficient is significant at the 10% level. as expected, this significant effect disappears in column 2 of Panel b for 2012 in the school-by-subject specification.

Table 13: **Estimated Effects of Weekly Instructional Hours and Class Periods on Achievement, by Classroom Quality (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Weekly Hours of Instruction	0.0269***	0.0104	0.0067	0.0121	0.0180***	-0.0025
	(0.0054)	(0.0082)	(0.0063)	(0.0078)	(0.0061)	(0.0076)
Weekly Hours × Overall Classroom Quality	0.0061	0.0061	0.0021	-0.0035	0.0121***	0.0040
	(0.0041)	(0.0064)	(0.0047)	(0.0069)	(0.0042)	(0.0066)
Panel (b): Weekly Number of Class Periods						
Weekly Number of Class Periods	0.0291***	0.0212**	0.0066	0.0323***	0.0168***	0.0408***
	(0.0053)	(0.0091)	(0.0054)	(0.0084)	(0.0054)	(0.0091)
Weekly Classes × Overall Classroom Quality	0.0067*	0.0044	0.0010	0.0013	0.0091**	0.0035
	(0.0038)	(0.0073)	(0.0039)	(0.0076)	(0.0035)	(0.0080)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	255,096	123,910	509,470	296,550	382,658	209,684
No. of Schools	13,190	5,412	10,706	5,316	12,290	5,596

Notes. Robust standard errors clustered at school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Consistent with the separate indexes, the 2015 combined interaction coefficients are all statistically insignificant across the school-by-grade and school-by-subject specification. However, it can be noted that the size of the coefficient increases for all results in the school-by-subject specification in 2015. In contrast, 2018 shows relatively significant positive interaction effects under the school-by-grade fixed effect model in column 1 that are larger in magnitude than even the student-related classroom quality instruction effects presented in Table 9. This suggests that in the 2018 wave, a comprehensive measure of classroom quality-capturing both teacher-related factors and student-related factors is a better predictor of instruction time effectiveness. As mentioned within the same section, this may because there is slightly more yearly variance in the distribution of responses for the year 2018 which explains the clearer, consistent results.

Table A13[30] broadens the analysis by including science test scores as well. Here we can observe that the strong positive interaction effects disappear and lose statistical

[30] See Page 53 in Appendix.

significance slightly. As mentioned before, although this may lead one to speculate that teacher-related behaviours are not important for science learning, other factors, such as measurement error in the science test scores, should be taken into account as well. After all, science was not the major test domain in 2018.

Table 14 explores non-linear effects by quartile. The result from Table 12 in 2012 disappears where the baseline group has the highest return for the school-by-grade specification. Instead, there is a significant interaction coefficient for both weekly hours of instruction and weekly number of class periods in the school-by-grade specification in the third combined classroom quality quartile. These effects, as expected, disappear in the school-by-subject specification.

Moving to 2015, the significant interaction effects disappear, mirroring the insignificant results observed in the linear specifications. In the school-by-grade specification, we find no significant interactions for either weekly hours or class periods. These effects remain largely absent in the school-by-subject specification, except for an isolated significant interaction in the second quartile for class periods. However, the 2018 results present a much clearer narrative of non-linear returns. In the school-by-grade specification, the baseline group shows no significant return to instruction time. Instead, the benefits are concentrated entirely at the top. For weekly hours, the interaction with the fourth quartile is 0.034 SD and is significant at the 1% level, resulting in a total effect size of 0.036 SD (0.002 + 0.034). This means that for students in the highest quartile of teacher-related classroom quality, an extra hour of weekly instruction is linked to a 0.036 SD increase in test scores for students while the effect for students in the lowest quartile is close to 0. Similarly, for class periods, the fourth quartile interaction is 0.022 SD and is significant at the 5% level, yielding a total effect of 0.026 SD (0.004 + 0.022). Consistent with the pattern observed in 2012, these significant interaction effects disappear in the school-by-subject specification, although the baseline effect for class periods in this specification becomes large (0.040) and significant at the 5% level.

Table 14: **Estimated Effects of Instructional Time on Achievement, by Quartile of Classroom Quality (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Instruction time	0.012	0.002	0.000	0.010	0.002	-0.006
	(0.009)	(0.013)	(0.011)	(0.013)	(0.010)	(0.012)
Instruction time × 2nd overall quality quartile	0.015	0.021	0.016	0.018	0.010	0.004
	(0.012)	(0.022)	(0.015)	(0.020)	(0.013)	(0.018)
Instruction time × 3rd overall quality quartile	0.029**	-0.005	0.014	-0.012	0.021*	-0.004
	(0.012)	(0.021)	(0.014)	(0.021)	(0.013)	(0.019)
Instruction time × 4th overall quality quartile	0.016	0.018	0.000	0.003	0.034***	0.011
	(0.012)	(0.021)	(0.014)	(0.020)	(0.013)	(0.021)
Panel (b): Weekly Number of Class Periods						
Instruction time	0.015	0.020	0.001	0.016	0.004	0.040**
	(0.009)	(0.015)	(0.009)	(0.014)	(0.008)	(0.016)
Instruction time × 2nd overall quality quartile	0.015	0.008	0.017	0.041**	0.010	-0.013
	(0.012)	(0.023)	(0.012)	(0.021)	(0.011)	(0.023)
Instruction time × 3rd overall quality quartile	0.027**	-0.018	0.009	0.009	0.018*	0.002
	(0.012)	(0.023)	(0.012)	(0.021)	(0.011)	(0.023)
Instruction time × 4th overall quality quartile	0.018	0.015	-0.001	0.018	0.022**	0.017
	(0.011)	(0.023)	(0.011)	(0.020)	(0.011)	(0.025)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	255,096	123,910	509,470	296,550	382,658	209,684
No. of Schools	13,190	5,412	10,706	5,316	12,290	5,596

Notes. Robust standard errors clustered at school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

When the analysis is expanded to include science scores in Table A14[31], none of the interaction effects are statistically significant at the top quartile of teacher-related classroom quality as observed in Table A12 as well. Again, while this may suggest that science performance is less closely tied to teacher behaviour and instructional support, these results should be interpreted cautiously because of the bias causing by the spillover effects between mathematics and science and reliability in the science domain that may affect the precision of the estimates.

7 Discussion of Results

The purpose was to move beyond baseline correlations and investigate whether the returns to additional hours are conditional on student and teacher-related factors. Therefore, this study was motivated by a critical research question: *How does the interaction between instruction time and student-related classroom quality influence student academic achievement across PISA 2012–2018?* The empirical analysis across the 2012, 2015, and 2018 cycles suggests that the answer is affirmative, yet complex. In this section, we aim to

[31] See Page 54 in Appendix.

interpret the varied support for instruction time and how it interacts with classroom quality (both student and teacher-related) observed across the years 2012-2018.

7.1 The Conditional Productivity of Time

Over the years observed in this study, we see empirical support for increasing instruction time with some caveats with the effect of the classroom learning environment. Our initial results show a positive relationship between these two factors. However, the magnitude of the effect size varies significantly by year and fixed effect specification. The drastic contrast between the robust results of the interaction effect in 2012 and 2018 and the insignificant results in 2015 implies that instruction time is an input that is easily impacted by methodological changes in PISA.

Furthermore, our findings challenge the assumption that added instruction time has a linear relationship with the academic achievement for students. Thus, we apply the Disruption Model in Lazear (2001), to show that the marginal benefit of an extra hour is largely determined by the probability of uninterrupted learning. Therefore, we identify that the interaction effect between an environment that is conducive to learning, named classroom quality in this paper, and instruction time is a vital component to impact academic results. We see this in our 2012 and 2018 data through the threshold effects observed, where significant gains are concentrated in the top quartile of classroom quality. This suggests that when classroom environment is poor, added instruction time has little impact on actually helping students learn more. In these lower-quartile environments, adding time results in diminishing returns, as the additional minutes are consumed by teachers trying to manage the classroom rather than academic instruction.

7.2 The PISA 2015 Anomaly

A distinct feature of our results is the divergence of results for the year 2015. In Table 7 and Figure 2, this is reflected in how the coefficient estimates for the school-by-grade and school-by-subject specifications change in their pattern of results. We attribute this primarily to the methodological change of the PISA survey going computer-based in most countries in 2015. This transition introduced a new form of error, digital literacy, into the dependent variable for the school-by-grade specification. This specification relies on within-student, cross-subject variation, this added noise of unfamiliarity with the testing interface likely causing measurement error. Additionally, the focus on science in the 2015 wave introduces potential precision issues. Due to the smaller amount of time allocated to language arts and mathematics test questions in the PISA survey, the test scores may fail to capture the direct impact of instructional hours, rendering the estimated effects less distinct. This may also be caused the the increased possibility of the language and mathematics test questions coming after the science section, which leads to student fatigue

as outlined in Section 6.2. The shift in significance and the break in pattern for the 2015 school-by-subject estimates in Table 7 may be caused by the larger sample size, which increases the ability to detect smaller effects.

To check if the change to the computer-based caused the unusual results, we attempted to analyse only the 15 countries that continued using paper tests in 2015. However, we could not generate reliable results because these countries had missing data regarding instruction time.[32]

Based on the return of significant results in 2018, which is consistent in most of our tables and figures but is perhaps most clearly observed in Table 7 and Figure 2, we speculate that this interference was temporary. We propose two mechanisms for this recovery:

1. Multi-stage adaptive testing introduced in 2018

   As mentioned in Section 4.2.6, in 2018, PISA introduced a multi-stage adaptive testing method for their major domain that year - language arts. This meant that students were given questions based on their ability, rather than being simply given questions in a linear, non-adaptive format like the previous cycles. This system prevented high-achieving students from topping out on easy questions, and low-ability students from guessing on questions that were too difficult. By attempting to ensure that every student faced a questionnaire appropriate to their level, 2018 likely captured a much more accurate picture of student ability at both the very top and very bottom of the distribution making results more precise overall. This explains why the school-by-subject specification gains statistical significance in Panel b of Table 7 for 2018 as learning persistence is captured to a certain degree. 10th-graders were likely to have received harder questions than 9th-graders in language arts due to learning persistence, and the general advantage of age and maturity due to the multi-adaptive testing structure.

2. The dissipation of the effects of digital testing

   In 2018, despite the test still being offered in a digital format, expected results are restored. This may be attributed to the fact that schools were more ready for the digital format in 2018 based on the previous cycle, as it was no longer new for the PISA test to be offered in a computer-based format as well. By 2018, digital familiarity had plausibly normalized across cohorts, reducing the magnitude of the ICT familiarity error term that confounded the 2015 data. According to Jerrim et al. (2018), the survey organisers recognise this issue and have implemented a strategy to lessen, though not entirely remove, the potential disruption. One approach used to address this issue was the introduction of multi-adaptive testing in PISA 2018, complemented

[32] We attempted a robustness check using only paper-based countries, but missing data on class period length for all 15 countries prevented us from conducting this analysis properly.

by the student-delivery-system manual, whose purpose is discussed in Section 4.2.6. Consequently, the 2018 results serve as a validation of our model, confirming that the odd-one-out patterns we see for 2015 results were a methodological outlier rather than a fundamental change in the relationship between instruction time and learning.

7.3 Interpreting the Teacher-Quality Extension

A goal of our thesis was to incorporate teacher-related factors alongside student-related ones, as both play a vital role in shaping a classroom environment conducive to learning. In 2012 and 2015, our teacher-related interaction terms were largely insignificant as can be seen in Table 11 and Table 12. We interpret this not as a lack of economic impact, but likely as a result of reporting bias. Due to the large variety of countries in our sample, schools in cultures that are less transparent or place a strong emphasis on protecting teachers' reputations may be less willing to report negative teacher qualities. This causes reporting bias. The lack of variance in responses to teacher-related factors that may stem from this can cause the small and statistically insignificant interaction estimates that we see for 2012 and 2015 in Table 11. When we compare supplementary Table A1 and A2, we observe that there is more year-to-year variation in responses to questions related to student behaviour than teacher behaviour. In fact, as shown in Table A2, the distribution of responses to teacher-related questions are almost identical for the year 2012 and 2015, with a very small proportion of schools (1%-4%) answering 'Not at all.'

However, our 2018 results regarding the combined classroom quality index, shown in Table 13, suggest that measuring student and teacher factors together provides a stronger predictor of achievement than examining them in isolation. The interaction effects in this combined model are larger and more significant than those found in the separate indexes (as seen in Table 9 and Table 11). This implies that a productive classroom environment depends on both student and teacher behaviour. Consequently, if either group falls short, whether through student-related factors or teacher-related factors, the benefit of additional instruction time appears to diminish. The combined classroom quality coefficient for weekly instruction hours is 0.0121 SD, which is higher than the teacher-related coefficient of 0.0102 SD and the student-related coefficient of 0.0118 SD. This indicates that improvements in overall classroom quality is correlated with larger gains in test scores per additional hour of instruction than improvements in teacher- or student-related factors alone. Furthermore, the combined classroom quality results are statistically significant at the 1% level for both weekly hours and weekly classes, whereas teacher-related factors were only significant at the 10% level.

7.4 Limitations of the Study

The findings in this paper should be interpreted with caution. Firstly, while our fixed effects strategy mitigates many endogeneity concerns, PISA data remains cross-sectional, which limits our ability to observe long-term causal trends. Furthermore, while the use of our fixed effect specification helps mitigate bias (Todd and Wolpin, 2003), it is important to keep in mind that the estimates do not show the true effect sizes and rather the 'ceiling' and 'floor' value within which the true effect size lies. Secondly, our measures of classroom quality rely on reports from school administrators, which may introduce systematic measurement error from cultural bias where specific countries always over-report quality. This would reduce the precision of our estimates and cause larger standard errors, which we see in our table, and reduce the variation available for your model to create the regressions from. Lastly, we measure the impact of the learning environment of the classroom, not the quality of the teacher (their ability, qualifications, experience) or the student (their innate ability, cumulative knowledge).

Finally, it is vital to look at these findings within the broader context of the educational environment. This study should not be viewed as a holistic evaluation of educational systems overall, but rather as a focused analysis of how the interaction between instruction time and classroom environment correlates with performance on standardised PISA tests. It cannot estimate how well a student will perform on any one test, or how to improve student scores for a specific curriculum, it can only explain what factors correlate with higher PISA scores for countries. There are various types of educational systems around the world, many with different objectives and goals which are not captured by standardised test scores. Thus, while this paper highlights the importance of efficiency in producing academic outcomes, the specific goals of different education systems mean that it might be necessary for them to make investments that are not immediately reflected in PISA test scores.

7.5 Economic Significance and Policy Implications

The findings of our study have implications for the design of educational policy and the efficient allocation of fiscal resources. Policy makers often view the extension of a school day as an efficient means to improve student outcomes (Radinger and Boeskens, 2021), assuming a linear production function where more time instruction time translates to a higher level of learning. However, our findings challenge this assumption, which suggest that an increase in instruction time without thorough judgment may be economically inefficient and, in contexts of low quality classroom environment (influenced both by student and teacher behaviour), may be fiscally inefficient.

Our analysis of diminishing returns to instruction time in Table 8 suggests that the marginal benefit of instruction time decreases as the total number of instruction hours in-

crease. All squared weekly hours of instruction coefficient estimates are negative and statistically significant at the 1% level, except in column 2 for 2015 under the school-by-grade specification. This suggests that policies requiring longer school hours, such as Florida's 2012 mandate for the lowest-performing reading schools to add an extra hour (Folsom et al., 2017) or the United Kingdom's 2022 initiative to increase weekly instructional time to 32.5 hours across all schools (Department of Education, 2023), may face a natural ceiling. From a budgetary perspective, increasing instruction time involves costs that include additional personnel hours and staffing arrangement (Kolbe and O'Reilly, 2016). If the marginal return to this time is diminishing as shown in Table 8, the cost per unit of increase in student achievement rises drastically at higher levels of instruction, suggesting that resources might be better allocated to other inputs once a certain threshold of time is reached.

Arguably, the most important economic finding of this study is the complementary relationship between instruction time and classroom environment. Our results in Section 6.3 and Section 6.4 imply that instruction time acts as a multiplier for classroom quality, rather than a substitute. Table 14 displays the estimated effects of instruction time (weekly hours and class periods) on achievement, by overall classroom quality (both student-related and teacher-related factors). In 2018, the total effect of an additional hour of instruction in the top quartile of combined classroom quality is 0.036 SD (0.002 + 0.034). In contrast, the effect for the lowest quartile is negligible. Likewise, the total effect of an additional class period in the top quartile was 0.026 SD (0.004 + 0.022) and the effect for the lowest quartile is statistically insignificant. Taken together, these results highlight that the benefits of additional instructional time depend heavily on classroom conditions, with high-quality environments amplifying learning gains and low-quality environments limiting them (Seidlitz and Zierow, 2025). Moreover, a key policy implication is that uniformly increasing instruction time across all schools may unintentionally widen achievement gaps, since students in higher-quality classrooms are better positioned to benefit from additional time. Higher-quality classroom are likely to be correlated with higher-quality schools in better neighborhoods, meaning that this could also widen the socioeconomic gap in achievement as well.

8 Conclusion

This thesis revisits the fundamental economic trade-off in education policy between the quantity of instruction time and the quality of the classroom environment. This study is inspired by the work of Rivkin and Schiman (2015), who analysed the 2009 PISA cycle for student-related class environment factors and their interaction effect with instruction time by employing rigorous school-by-grade and school-by-subject fixed effects. We have expanded the analysis to cover the 2012, 2015, and 2018 PISA cycles, as well as included

teacher-related classroom environment which gives us a more holistic view of the classroom environment. We have attempted to isolate the causal impact of instruction time while controlling for non-random sorting and unobserved heterogeneity.

We find that while additional instruction time is positively correlated with academic achievement, it is subject to diminishing returns at high hours. More importantly, we provide evidence that the marginal productivity of instruction time is highly heterogeneous and dependent on the classroom environment. For student-related classroom quality we find that instruction time and classroom quality are complementary. However, the evidence for teacher-related classroom quality is much less significant and shows smaller effect sizes. The combined classroom quality index confirms that instruction time is complementary to the overall classroom environment. This means that you cannot simply add hours to a school day with poor overall quality and expect results to increase drastically, the classroom environment needs to be improved to the third or fourth quality quartile for the additional time to translate into higher academic achievement. Consistent with the Disruption Model in Lazear (2001), we find that additional hours are most effective in environments characterized by low student disruption and strong teacher support. Furthermore, our findings are largely consistent to the pattern seen by Rivkin and Schiman (2015). The results from 2018, highlight that a comprehensive measure of quality combining both student and teacher behaviour is a stronger predictor of time effectiveness than looking at student factors in isolation. This suggests that the return on investment for additional school hours is dependent on whether or not the environment in the classroom is conducive to learning.

These findings carry significant implications for education policy and budget allocation. Policy makers often view extending the school day as the obvious method to improve school outcomes. However, our analysis suggests that such policies may be fiscally inefficient if implemented in isolation. If the classroom environment experiences high levels of disruption or frequent teacher absenteeism, simply mandating more hours will lead to minimal gains. Therefore, resources should be prioritised toward interventions that improve the overall quality of the learning environment, such as lowering teacher absenteeism and reducing bullying between students, alongside investments in extending instruction time in order to increase school results.

Looking forward, it would be valuable for future research to investigate teacher-related classroom quality in a broader manner that incorporates teaching qualifications and teacher experience, in addition to teacher behaviour. Similarly, future research could examine whether the interaction between classroom quality and instruction hours varies by subjects. Lastly, although we exclude 2022 because it is not comparable to earlier years due to the exogenous shock of COVID-19, it would be useful for future research to examine how pandemic-related developments such as online schooling and increased ICT use in classrooms influenced the relationship between classroom quality, instructional hours, and student achievement.

References

Bray, M. (2021). Shadow education in Europe: Growing prevalence, underlying forces, and policy implications. *ECNU Review of Education*, *4*(3), 442–475. https://doi.org/10.1177/2096531119890142

Britannica. (n.d.). Cell phones in schools. https://www.britannica.com/procon/cell-phones-debate

Chauhan, S. (2017). A meta-analysis of the impact of technology on learning effectiveness of elementary students. *Computers & Education*, *105*, 14–30. https://doi.org/10.1016/J.COMPEDU.2016.11.005

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/QJE/QJR041

Council of Ministers of Education Canada. (n.d.). *Programme for International Student Assessment (PISA): Overview*. https://www.cmec.ca/251/Programme_for_International_Student_Assessment_(PISA).html

Dee, T. S., & Cohodes, S. R. (2008). Out-of-field teachers and student achievement. *Public Finance Review*, *36*(1), 7–32. https://doi.org/10.1177/1091142106289330

Department for Education. (2023). Length of the school week: Non-statutory guidance. https://assets.publishing.service.gov.uk/media/64a2f21fbb13dc000cb2e5e1/Minimum_School_Week_Non-Statutory_Guidance.pdf

European Commission. (2019). Integrating students from migrant backgrounds into schools in Europe: National policies and measures. *Publications Office of the European Union.* https://eurydice.eacea.ec.europa.eu/publications/integrating-students-migrant-backgrounds-schools-europe-national-policies-and-measures

Folsom, J. S., Osborne-Lampkin, L., Cooley, S., & Smith, K. (2017). Implementing the extended school day policy in Florida's 300 lowest performing elementary schools. *Institute of Education Sciences.* https://ies.ed.gov/use-work/resource-library/report/descriptive-study/implementing-extended-school-day-policy-floridas-300-lowest-performing-elementary-schools

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education*, *44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Kolbe, T., & O'Reilly, F. (2017). The cost of increasing in-school time: Evidence from the Massachusetts Expanded Learning Time Initiative. *Leadership and Policy in Schools*, *16*(4), 563–601. https://doi.org/10.1080/15700763.2016.1232832

Komatsu, H., & Rappleye, J. (2017). Did the shift to computer-based testing in PISA 2015 affect reading scores? A view from East Asia. *Compare: A Journal of Comparative and International Education*, *47*(4), 616–623. https://doi.org/10.1080/03057925.2017.1309864

Lavy, V. (2010). Effects of free choice among public schools. *The Review of Economic Studies*, *77*(3), 1164–1191. https://doi.org/10.1111/J.1467-937X.2009.00588.X

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal*, *125*(588), F397–F424. https://www.jstor.org/stable/24737554

Lavy, V. (2020). Expanding school resources and increasing time on task: Effects on students' academic and noncognitive outcomes. *Journal of the European Economic Association*, *18*(1), 232–265. https://doi.org/10.1093/JEEA/JVY054

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, *116*(3), 777–803. https://www.jstor.org/stable/2696418

Meehan, L., Pacheco, G., & Schober, T. (2023). Basic reading and mathematics skills and the labour market outcomes of young people: Evidence from PISA and linked administrative data. *The Economic Record*, *99*(327), 473–491. https://doi.org/10.1111/1475-4932.12755

Meghir, C., & Rivkin, S. (2011). Econometric methods for research in education. *Handbook of the Economics of Education*, *3*, 1–87. https://doi.org/10.1016/B978-0-444-53429-3.00001-6

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy and Practice*, *26*(4), 422–443. https://doi.org/10.1080/0969594X.2018.1449100

National Center for Education Statistics. (n.d.). *How does the PISA multi-stage adaptive test of reading work?* https://nces.ed.gov/surveys/pisa/pisa2018/index.asp

Niki, M. (2024). Does the reduction in instruction time affect student achievement and motivation? Evidence from Japan. *Japan and the World Economy*, *70*, 101254. https://doi.org/10.1016/J.JAPWOR.2024.101254

OECD. (n.d.). PISA: Programme for International Student Assessment. *OECD Publishing*. https://www.oecd.org/en/about/programmes/pisa.html

OECD. (2009). PISA data analysis manual: SPSS (2nd ed.). *OECD Publishing*.https://www.oecd.org/content/dam/oecd/en/publications/reports/2009/03/pisa-data-analysis-manual-spss-second-edition

OECD. (2012). PISA 2012 school questionnaire. *OECD Publishing*. https://www.oecd.org/content/dam/oecd/en/data/datasets/pisa/pisa-2012-datasets/main-survey/questionnaires/School%20questionnaire%20PISA%202012.pdf

OECD. (2016). PISA 2018 student delivery system manual (field trial). *OECD Publishing*. https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/pisa-database/survey-implementation-tools/pisa-2018/STUDENT-DELIVERY-SYSTEM-MANUAL-Field-Trial.pdf

OECD. (2023a). PISA 2022 assessment and analytical framework. *OECD Publishing*. https://doi.org/10.1787/dfe0bf9c-en

OECD. (2023b). PISA 2022 results (Volume I): The state of learning and equity in education. *OECD Publishing*. https://doi.org/10.1787/53f23881-en

OECD. (2024). PISA 2022 technical report. *OECD Publishing*. https://doi.org/10.1787/01820D6D-EN

Okubo, T. (2022). Theoretical considerations on scaling methodology in PISA (OECD Education Working Papers No. 282). *OECD Publishing*. https://doi.org/10.1787/c224dbeb-en

Perkins, R., & Shiel, G. (2016). A teacher's guide to PISA mathematics and problem solving: Findings from PISA 2012. *Educational Research Centre*. http://www.erc.ie

Radinger, T., & Boeskens, L. (2021). More time at school: Lessons from case studies and research on extended school days (OECD Education Working Papers No. 252). *OECD Publishing*. https://doi.org/10.1787/1f50c70d-en

Rivkin, S. G., & Schiman, J. C. (2015). Instruction time, classroom quality, and academic achievement. *The Economic Journal*, *125*(588), F425–F448. https://doi.org/10.1111/ECOJ.12315

Robitzsch, A., Lüdtke, O., Goldhammer, F., Kroehne, U., & Köller, O. (2020). Reanalysis of the German PISA data: A comparison of different approaches for trend estimation with a particular emphasis on mode effects. *Frontiers in Psychology*, *11*, 528385. https://doi.org/10.3389/FPSYG.2020.00884

Schuelka, M. J. (2013). Excluding students with disabilities from the culture of achievement: The case of the TIMSS, PIRLS, and PISA. *Journal of Education Policy*, *28*(2), 216–230. https://doi.org/10.1080/02680939.2012.708789

Seidlitz, A., & Zierow, L. (2025). Longer days, better performance? The impact of all-day primary schools in Germany. *Economics of Education Review*, *107*, 102659. https://doi.org/10.1016/J.ECONEDUREV.2025.102659

Sjøberg, S. (2015). PISA and global educational governance - A critique of the project, its uses and implications. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(1), 111–127. https://doi.org/10.12973/EURASIA.2015.1310A

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33. https://doi.org/10.1111/1468-0297.00097

Wedel, K. (2021). Instruction time and student achievement: The moderating role of teacher qualifications. *Economics of Education Review*, *85*, 102183. https://doi.org/10.1016/J.ECONEDUREV.2021.102183

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/J.STUEDUC.2005.05.005

National Center for Education Statistics. (n.d.). OECD Program for International Student Assessment 2012 School Questionnaire. https://nces.ed.gov/surveys/pisa/pdf/MS12_ScQ_USA_final.pdf

A Appendix

Table A1: **Distribution of Responses to Student-Related Classroom Quality Measures for PISA 2012-2018**

Table A1.1: PISA 2012

	'In your school, to what extent is the learning of students hindered by the following phenomenon?'				
	Not at all	Very little	To some extent	A lot	Total
Student truancy	0.09	0.28	0.44	0.19	17,470
Students skip classes	0.07	0.26	0.48	0.19	17,468
Students lack respect	0.03	0.16	0.52	0.28	17,482
Student alcohol/drug use	0.03	0.06	0.33	0.59	17,424
Student bullying	0.02	0.10	0.53	0.35	17,268

Table A1.2: PISA 2015

	'In your school, to what extent is the learning of students hindered by the following phenomenon?'				
	Not at all	Very little	To some extent	A lot	Total
Student truancy	0.08	0.28	0.46	0.17	16,319
Students skip classes	0.06	0.26	0.48	0.19	16,289
Students lack respect	0.04	0.18	0.54	0.24	16,278
Student alcohol/drug use	0.03	0.07	0.36	0.54	16,269
Student bullying	0.02	0.11	0.56	0.31	16,288

Table A1.3: PISA 2018

	'In your school, to what extent is the learning of students hindered by the following phenomenon?'				
	Not at all	Very little	To some extent	A lot	Total
Student truancy	0.12	0.29	0.41	0.18	20,823
Students skip classes	0.09	0.28	0.45	0.18	20,822
Students lack respect	0.06	0.19	0.50	0.25	20,800
Student alcohol/drug use	0.07	0.07	0.33	0.53	20,811
Student bullying	0.05	0.11	0.52	0.32	20,823

Notes. **Data source:** The OECD and PISA.

Table A2: **Distribution of Responses to Teacher-Related Classroom Quality Measures for PISA 2012-2018**

Table A2.1: PISA 2012

	'In your school, to what extent is the learning of students hindered by the following phenomenon?'				
	Not at all	Very little	To some extent	A lot	Total
Teachers too strict with students	0.01	0.12	0.54	0.33	17,255
Teachers not meeting students needs	0.03	0.21	0.49	0.27	17,257
Teacher absenteeism	0.04	0.12	0.46	0.39	17,267
Staff resisting change	0.04	0.23	0.44	0.29	17,232
Teachers not well-prepared	0.03	0.09	0.48	0.40	17,231

Table A2.2: PISA 2015

	'In your school, to what extent is the learning of students hindered by the following phenomenon?'				
	Not at all	Very little	To some extent	A lot	Total
Teachers too strict with students	0.01	0.14	0.54	0.32	16,278
Teachers not meeting students needs	0.02	0.20	0.49	0.29	16,283
Teacher absenteeism	0.03	0.14	0.46	0.37	16,320
Staff resisting change	0.04	0.24	0.43	0.29	16,270
Teachers not well-prepared	0.03	0.11	0.50	0.35	16,284

Table A2.3: PISA 2018

	'In your school, to what extent is the learning of students hindered by the following phenomenon?'				
	Not at all	Very little	To some extent	A lot	Total
Teachers too strict with students	0.03	0.15	0.52	0.31	20,832
Teachers not meeting students needs	0.05	0.25	0.49	0.22	20,820
Teacher absenteeism	0.06	0.14	0.43	0.37	20,818
Staff resisting change	0.06	0.24	0.42	0.29	20,817
Teachers not well-prepared	0.07	0.12	0.47	0.34	20,825

Notes. **Data source:** The OECD and PISA.

Table A3: **Distributions of Weekly Language Arts Instructional Minutes and Classes Conditional on Mathematics Instruction Time for (2012- 2018)**

Table A3.1: PISA 2012

Panel (a): Minutes per week

	Mathematics							
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+	Observations
0–99	0.434	0.050	0.024	0.010	0.024	0.011	0.016	5,116
100–179	0.383	0.562	0.158	0.231	0.120	0.095	0.069	26,690
180–199	0.084	0.132	0.498	0.011	0.181	0.100	0.037	19,881
200–219	0.023	0.110	0.007	0.604	0.008	0.155	0.056	17,473
220–239	0.018	0.075	0.138	0.006	0.520	0.069	0.031	15,462
240–279	0.017	0.041	0.117	0.085	0.089	0.450	0.173	22,417
280+	0.040	0.030	0.057	0.053	0.057	0.120	0.617	21,469
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Observations	5,480	20,860	20,091	15,212	15,993	27,407	23,465	128,508

Panel (b): Number of class periods per week

	Mathematics					
Language arts	0–2	3	4	5	6+	Observations
0–2	0.544	0.114	0.029	0.033	0.043	10,966
3	0.287	0.495	0.180	0.081	0.071	25,828
4	0.106	0.296	0.583	0.289	0.137	44,660
5	0.045	0.059	0.126	0.509	0.165	28,042
6+	0.019	0.036	0.083	0.089	0.583	19,012
Total	1.000	1.000	1.000	1.000	1.000	
Observations	9,446	23,175	41,131	35,063	19,693	128,508

Table A3.2: PISA 2015

Panel (a): Minutes per week

	Mathematics							
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+	Observations
0–99	0.472	0.063	0.031	0.022	0.046	0.024	0.032	16,713
100–179	0.314	0.594	0.165	0.265	0.108	0.110	0.072	56,651
180–199	0.091	0.149	0.556	0.002	0.212	0.101	0.031	43,926
200–219	0.015	0.090	0.002	0.577	0.000	0.172	0.049	29,944
220–239	0.021	0.057	0.149	0.000	0.486	0.065	0.032	29,912
240–279	0.017	0.032	0.073	0.093	0.081	0.434	0.143	38,978
280+	0.069	0.015	0.024	0.041	0.066	0.094	0.641	41,308
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Observations	16,838	42,786	39,783	25,088	33,549	52,002	47,386	257,432

Panel (b): Number of class periods per week

	Mathematics					
Language arts	0–2	3	4	5	6+	Observations
0–2	0.510	0.111	0.034	0.056	0.059	25,595
3	0.280	0.516	0.184	0.080	0.075	50,577
4	0.122	0.293	0.589	0.275	0.133	83,016
5	0.064	0.061	0.144	0.496	0.184	58,023
6+	0.023	0.019	0.049	0.092	0.550	40,221
Total	1.000	1.000	1.000	1.000	1.000	
Observations	22,595	41,994	71,500	68,436	52,907	257,432

Table A3.3: PISA 2018

Panel (a): Minutes per week

	Mathematics							
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+	Observations
0–99	0.487	0.061	0.031	0.021	0.032	0.016	0.027	11,021
100–179	0.289	0.592	0.144	0.248	0.115	0.103	0.071	39,647
180–199	0.084	0.139	0.560	0.002	0.155	0.100	0.037	29,507
200–219	0.015	0.097	0.002	0.620	0.000	0.169	0.052	25,067
220–239	0.023	0.056	0.158	0.000	0.583	0.066	0.053	27,561
240–279	0.023	0.035	0.077	0.073	0.062	0.455	0.128	27,818
280+	0.079	0.019	0.027	0.037	0.053	0.091	0.632	31,671
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
Observations	11,159	29,087	26,747	22,588	29,356	35,568	37,787	192,292

Panel (b): Number of class periods per week

	Mathematics					
Language arts	0–2	3	4	5	6+	Observations
0–2	0.531	0.112	0.030	0.060	0.051	18,601
3	0.253	0.532	0.170	0.088	0.076	37,704
4	0.117	0.273	0.636	0.255	0.124	66,185
5	0.069	0.061	0.121	0.510	0.243	43,511
6+	0.031	0.022	0.042	0.087	0.506	26,291
Total	1.000	1.000	1.000	1.000	1.000	
Observations	16,269	30,303	61,861	47,553	36,306	192,292

Notes. **Data source:** The OECD and PISA.

Table A4: **Average Mathematics Minus Language Arts Standardised Score Difference by Instructional Time in Mathematics and Language Arts for PISA 2012-2018**

Table A4.1: PISA 2012

Panel (a): minutes per week

	Mathematics						
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+
0–99	-0.020	-0.005	-0.122	0.015	0.037	0.050	-0.034
100–179	-0.031	-0.035	0.018	-0.040	0.009	-0.025	0.040
180–199	-0.060	0.013	0.001	-0.057	-0.019	0.074	0.033
200–219	-0.284	-0.053	0.047	0.003	-0.030	-0.016	0.045
220–239	-0.049	-0.068	-0.007	-0.028	0.044	0.008	0.054
240–279	-0.022	-0.034	-0.139	0.031	0.111	0.001	0.026
280+	-0.051	-0.136	-0.134	0.036	0.071	0.099	0.011

Panel (b): number of class periods per week

	Mathematics				
Language arts	0–2	3	4	5	6+
0–2	-0.035	0.015	-0.080	-0.018	0.051
3	-0.024	-0.037	0.015	-0.013	0.066
4	-0.052	-0.073	0.028	-0.019	0.049
5	-0.125	-0.113	-0.012	-0.037	-0.049
6+	-0.068	-0.202	0.071	0.110	0.102

Table A4.2: PISA 2015

Panel (a): minutes per week

	Mathematics						
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+
0–99	0.009	0.021	-0.016	-0.013	-0.044	0.013	-0.062
100–179	0.036	-0.030	-0.047	-0.043	-0.030	0.008	0.114
180–199	0.002	0.057	-0.024	0.106	0.002	0.001	0.082
200–219	0.062	0.056	0.352	-0.020	-0.256	0.011	0.149
220–239	0.052	-0.042	0.069	0.646	0.001	-0.002	0.068
240–279	-0.047	0.049	-0.083	0.071	0.061	-0.008	0.055
280+	-0.110	0.007	-0.017	0.112	0.116	0.093	-0.032

Panel (b): number of class periods per week

	Mathematics				
Language arts	0–2	3	4	5	6+
0–2	0.003	-0.014	0.009	-0.042	0.079
3	0.025	-0.043	-0.032	0.004	0.023
4	0.010	0.012	-0.024	0.006	0.040
5	-0.053	-0.005	0.049	-0.056	-0.006
6+	0.004	-0.059	0.068	0.087	0.078

Table A4.3: PISA 2018

Panel (a): minutes per week

	Mathematics						
Language arts	0–99	100–179	180–199	200–219	220–239	240–279	280+
0–99	-0.050	-0.023	-0.035	-0.005	-0.088	0.017	-0.126
100–179	0.019	0.025	-0.042	0.057	-0.002	0.094	0.278
180–199	-0.052	0.009	-0.017	-0.014	-0.038	-0.016	0.041
200–219	0.056	0.093	0.222	0.016	0.012	-0.015	0.134
220–239	0.052	0.001	0.014		0.007	-0.010	0.007
240–279	0.065	0.058	-0.031	0.082	-0.001	-0.046	-0.016
280+	-0.067	0.056	-0.067	0.128	0.077	0.078	-0.056

Panel (b): number of class periods per week

	Mathematics				
Language arts	0–2	3	4	5	6+
0–2	-0.036	-0.006	0.017	-0.026	0.274
3	0.016	-0.010	0.007	0.073	0.095
4	0.029	0.027	0.011	-0.000	0.046
5	-0.040	0.023	0.013	-0.080	-0.037
6+	0.140	-0.016	0.039	0.043	0.015

Notes. Reported values are average differences (Math − Language Arts). Positive values indicate Math performance exceeds Language Arts. **Data source:** The OECD and PISA.

Table A5: **Percent of Variation in Instruction Time Measures Explained by the Fixed Effects (2012-2018)**

	(1)	(2)	(3)	(4)	(5)
Panel (a): Weekly Hours of Instruction					
2012	0%	54%	91%	89%	98%
2015	0%	61%	93%	92%	99%
2018	0%	55%	92%	87%	98%
Panel (b): Weekly Number of Class Periods					
2012	0%	55%	91%	92%	98%
2015	0%	59%	92%	94%	99%
2018	0%	54%	91%	92%	98%
Fixed Effects					
Country-by-grade-by-subject	N	Y	Y	Y	Y
School-by-grade	N	N	Y	N	Y
School-by-subject	N	N	N	Y	Y

Notes. The table reports the R^2 values, representing the proportion of variation in class frequency and length explained by different fixed effects (FE). Column (1) presents the baseline variation without controls. Column (2) accounts for national and grade-level standardisations. Columns (3) and (4) account for policies specific to a school-grade or school-subject, respectively. Column (5) accounts for both school-grade and school-subject factors simultaneously. **Data source:** The OECD and PISA.

Table A6: **Estimated Effects of Weekly Hours of Instruction and Class Periods on Achievement in Mathematics, Science, and Language Arts (2012-2018)**

	2012			2015			2018		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Panel (a): Weekly Hours of Instruction									
Weekly Hours of Instruction	0.125***	0.009***	0.029***	0.165***	0.017***	0.028***	0.125***	0.009***	0.029***
	(0.007)	(0.003)	(0.006)	(0.008)	(0.002)	(0.006)	(0.007)	(0.003)	(0.006)
Panel (b): Weekly Number of Class Periods									
Weekly Number of Class Periods	0.144***	0.008***	0.040***	0.182***	0.015***	0.042***	0.144***	0.008***	0.040***
	(0.006)	(0.002)	(0.006)	(0.008)	(0.002)	(0.006)	(0.006)	(0.002)	(0.006)
School-by-grade FE	N	Y	N	N	Y	N	N	Y	N
School-by-subject FE	N	N	Y	N	N	Y	N	N	Y
Sample Size	478,890	478,890	238,857	745,629	745,629	428,355	478,890	478,890	238,857
No. of Schools	13,344	13,344	5,748	10,798	10,798	5,338	13,344	13,344	5,748

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Table A7: **Out-of-school study for PISA 2012 and 2015**

	2012				2015			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Panel (a): Weekly Hours of Instruction								
Instruction time	0.027***	0.026***	0.006	0.007	0.048***	0.050***	0.023	0.033**
	(0.005)	(0.006)	(0.008)	(0.008)	(0.011)	(0.011)	(0.016)	(0.016)
Out-of-school lessons (rel. to 2 to <4 hours)								
None		0.062**		0.305***		0.051		0.257***
		(0.024)		(0.033)		(0.038)		(0.048)
<2 hours per week		-0.050**		-0.026		0.026*		0.141***
		(0.024)		(0.037)		(0.014)		(0.021)
4 to <6 hours per week		0.087***		-0.045		-0.009		-0.103***
		(0.033)		(0.046)		(0.017)		(0.025)
6+ hours per week		0.128**		-0.054		0.033		-0.092***
		(0.050)		(0.062)		(0.035)		(0.035)
Panel (b): Weekly Number of Class Periods								
Instruction time	0.030***	0.029***	0.018**	0.020**	0.037***	0.040***	0.042***	0.050***
	(0.005)	(0.005)	(0.009)	(0.009)	(0.008)	(0.009)	(0.016)	(0.015)
Out-of-school lessons (rel. to 2 to <4 hours)								
None		0.066***		0.305***		0.050		0.257***
		(0.024)		(0.033)		(0.038)		(0.048)
<2 hours per week		-0.047*		-0.026		0.026*		0.142***
		(0.024)		(0.037)		(0.014)		(0.021)
4 to <6 hours per week		0.085***		-0.047		-0.009		-0.103***
		(0.033)		(0.046)		(0.017)		(0.025)
6+ hours per week		0.123**		-0.061		0.034		-0.092***
		(0.050)		(0.062)		(0.035)		(0.035)
School-by-grade FE	Y	Y	N	N	Y	Y	N	N
School-by-subject FE	N	N	Y	Y	N	N	Y	Y
Sample Size	255,864		124,067		206,616		127,382	
No. of schools	13,242		5,435		4,550		2,425	

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Results limited to 2012 and 2015, as 2018 lacks comparable out-of-school time data. The 2015 sample represents only the 21 countries that administered the optional educational career questionnaire out of 49-country sample. **Data source:** The OECD and PISA.

Table A8: **Estimated Effects of Weekly Instructional Minutes and Classes per Week on Achievement from Specifications with Instructional Time Categories (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Minutes per week (relative to 200–219)						
0–99	-0.100**	-0.283***	-0.050	-0.343***	-0.030	-0.386***
	(0.039)	(0.057)	(0.036)	(0.058)	(0.035)	(0.062)
100–179	-0.097***	-0.026	-0.063***	-0.116***	-0.029	-0.102***
	(0.021)	(0.034)	(0.024)	(0.038)	(0.022)	(0.039)
180–199	-0.112***	-0.026	-0.056**	-0.020	0.007	-0.067
	(0.025)	(0.043)	(0.027)	(0.045)	(0.025)	(0.047)
220–239	-0.017	0.038	-0.039	0.011	0.019	0.071
	(0.026)	(0.045)	(0.029)	(0.048)	(0.027)	(0.047)
240–279	0.059***	0.014	-0.018	0.027	0.048*	0.003
	(0.021)	(0.038)	(0.024)	(0.040)	(0.025)	(0.042)
280+	0.023	-0.060	-0.018	-0.069	0.034	-0.123***
	(0.025)	(0.043)	(0.030)	(0.045)	(0.028)	(0.046)
Panel (b): Classes per week (relative to 4)						
0–2	-0.090***	-0.182***	-0.016	-0.223***	-0.061***	-0.193***
	(0.025)	(0.042)	(0.024)	(0.044)	(0.024)	(0.046)
3	-0.129***	-0.017	-0.058***	-0.048*	-0.024	-0.005
	(0.014)	(0.024)	(0.016)	(0.026)	(0.014)	(0.032)
5	0.024*	0.017	-0.010	0.037	0.002	0.059**
	(0.014)	(0.030)	(0.015)	(0.028)	(0.016)	(0.030)
6–10	0.011	-0.004	-0.009	-0.025	0.053***	0.089**
	(0.019)	(0.039)	(0.022)	(0.037)	(0.021)	(0.040)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	256,316	124,456	513,778	299,312	384,302	210,596
No. of Schools	13,263	5,439	10,801	5,369	12,349	5,624

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Table A9: **Estimated Effects of Weekly Instructional Hours and Class Periods by Student-Related Classroom Quality in Mathematics, Science, and Language Arts (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Weekly Hours of Instruction	0.0083***	0.0300***	0.0175***	0.0287***	0.0117***	0.0097
	(0.0026)	(0.0064)	(0.0023)	(0.0063)	(0.0025)	(0.0059)
Weekly Hours × Student-related Classroom Quality	0.0062***	0.0022	0.0008	0.0057	0.0052**	0.0091
	(0.0020)	(0.0053)	(0.0018)	(0.0056)	(0.0021)	(0.0059)
Panel (b): Weekly Number of Class Periods						
Weekly Number of Class Periods	0.0081***	0.0404***	0.0148***	0.0430***	0.0108***	0.0471***
	(0.0023)	(0.0062)	(0.0019)	(0.0059)	(0.0021)	(0.0062)
Weekly Classes × Student-related Classroom Quality	0.0053***	0.0054	0.0000	0.0086*	0.0038**	0.0072
	(0.0017)	(0.0055)	(0.0015)	(0.0051)	(0.0018)	(0.0059)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	480,132	239,652	745,629	428,355	683,151	379,068
No. of Schools	13,387	5,769	10,798	5,338	12,446	5,876

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Table A10: **Estimated Effects of Instructional Time by Quartile of Student-Related Classroom Quality in Mathematics, Science, and Language Arts (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Instructional Hours						
Instruction time	0.002	0.016	0.014***	0.017	0.008**	0.009
	(0.004)	(0.012)	(0.004)	(0.011)	(0.004)	(0.010)
Instruction time × 2nd student-related quality quartile	0.002	0.020	0.004	0.020	0.002	-0.002
	(0.005)	(0.017)	(0.005)	(0.015)	(0.005)	(0.014)
Instruction time × 3rd student-related quality quartile	0.008	0.030*	0.010*	0.014	0.005	-0.007
	(0.005)	(0.017)	(0.005)	(0.016)	(0.005)	(0.015)
Instruction time × 4th student-related quality quartile	0.016***	0.011	0.002	0.011	0.009	0.014
	(0.005)	(0.016)	(0.005)	(0.017)	(0.005)	(0.016)
Panel (b): Weekly Number of Class Periods						
Instruction time	0.003	0.025**	0.013***	0.023**	0.009**	0.044***
	(0.004)	(0.012)	(0.003)	(0.010)	(0.003)	(0.011)
Instruction time × 2nd student-related quality quartile	0.001	0.017	0.003	0.034***	0.002	-0.007
	(0.005)	(0.017)	(0.004)	(0.013)	(0.004)	(0.015)
Instruction time × 3rd student-related quality quartile	0.006	0.028*	0.006	0.023*	0.003	0.008
	(0.005)	(0.016)	(0.004)	(0.014)	(0.004)	(0.015)
Instruction time × 4th student-related quality quartile	0.013***	0.017	-0.000	0.021	0.004	0.015
	(0.005)	(0.016)	(0.004)	(0.015)	(0.005)	(0.016)
School-by-grade FE	Y	N	Y	N	Y	N
School-by-subject FE	N	Y	N	Y	N	Y
Sample Size	480,132	239,652	747,201	428,622	683,151	379,068
No. of Schools	13,387	5,769	10,828	5,343	12,446	5,876

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Table A11: **Estimated Effects of Weekly Instructional Hours and Class Periods by Teacher-Related Classroom Quality in Mathematics, Science, and Language Arts (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Weekly Hours of Instruction	0.0083***	0.0314***	0.0176***	0.0280***	0.0122***	0.0109*
	(0.0026)	(0.0065)	(0.0023)	(0.0064)	(0.0025)	(0.0060)
Weekly Hours × Classroom Quality	-0.0017	0.0083	0.0009	0.0011	0.0022	0.0061
	(0.0020)	(0.0056)	(0.0017)	(0.0057)	(0.0020)	(0.0058)
Panel (b): Weekly Number of Class Periods						
Weekly Number of Class Periods	0.0081***	0.0404***	0.0081***	0.0404***	0.0081***	0.0404***
	(0.0023)	(0.0062)	(0.0023)	(0.0062)	(0.0023)	(0.0062)
Weekly Classes × Classroom Quality	0.0053***	0.0054	0.0053***	0.0054	0.0053***	0.0054
	(0.0017)	(0.0055)	(0.0017)	(0.0055)	(0.0017)	(0.0055)
School-by-grade fixed effect	Y	N	Y	N	Y	N
School-by-subject fixed effect	N	Y	N	Y	N	Y
Sample Size	481,098	240,180	748,734	428,919	684,525	379,182
Number of Schools	13,407	5,780	10,845	5,352	12,471	5,882

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Table A12: **Estimated Effects of Instructional Time by Quartile of Teacher-Related Classroom Quality in Mathematics, Science, and Language Arts (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Instruction time	0.011***	0.017	0.012***	0.032***	0.008**	0.004
	(0.004)	(0.011)	(0.004)	(0.010)	(0.004)	(0.009)
Instruction time × 2nd quality quartile	-0.004	0.013	0.010**	-0.012	0.007	0.002
	(0.005)	(0.016)	(0.005)	(0.015)	(0.005)	(0.014)
Instruction time × 3rd quality quartile	0.000	0.024	0.009*	-0.006	0.003	0.027*
	(0.005)	(0.016)	(0.005)	(0.015)	(0.005)	(0.016)
Instruction time × 4th quality quartile	-0.007	0.022	0.003	-0.001	0.005	0.005
	(0.005)	(0.017)	(0.005)	(0.015)	(0.005)	(0.015)
Panel (b): Weekly Number of Class Periods						
Instruction time	0.009***	0.030***	0.012***	0.040***	0.007**	0.045***
	(0.004)	(0.011)	(0.003)	(0.009)	(0.003)	(0.010)
Instruction time × 2nd quality quartile	-0.002	0.016	0.008*	-0.002	0.008*	0.002
	(0.005)	(0.015)	(0.004)	(0.014)	(0.004)	(0.014)
Instruction time × 3rd quality quartile	0.001	0.020	0.006*	-0.004	0.004	0.009
	(0.005)	(0.016)	(0.004)	(0.013)	(0.004)	(0.015)
Instruction time × 4th quality quartile	-0.004	0.011	0.001	0.010	0.003	0.003
	(0.005)	(0.016)	(0.004)	(0.014)	(0.004)	(0.015)
School-by-grade fixed effect	Y	N	Y	N	Y	N
School-by-subject fixed effect	N	Y	N	Y	N	Y
Sample Size	481,098	240,180	748,734	428,919	684,525	379,182
Number of Schools	13,407	5,780	10,845	5,352	12,471	5,882

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Table A13: **Estimated Effects of Weekly Instructional Hours and Class Periods by Classroom Quality in Mathematics, Science, and Language Arts (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Weekly Hours of Instruction	0.0085***	0.0310***	0.0176***	0.0285***	0.0121***	0.0113*
	(0.0026)	(0.0064)	(0.0023)	(0.0064)	(0.0025)	(0.0060)
Weekly Hours × Classroom Quality	0.0028	0.0047	0.0006	0.0039	0.0046**	0.0084
	(0.0020)	(0.0055)	(0.0018)	(0.0057)	(0.0021)	(0.0060)
Panel (b): Weekly Number of Class Periods						
Weekly Number of Class Periods	0.0082***	0.0413***	0.0149***	0.0432***	0.0110***	0.0476***
	(0.0023)	(0.0062)	(0.0020)	(0.0060)	(0.0021)	(0.0062)
Weekly Classes × Classroom Quality	0.0028	0.0034	-0.0001	0.0070	0.0033*	0.0052
	(0.0017)	(0.0056)	(0.0015)	(0.0054)	(0.0018)	(0.0060)
School-by-grade fixed effect	Y	N	Y	N	Y	N
School-by-subject fixed effect	N	Y	N	Y	N	Y
Sample Size	476,664	237,798	739,401	424,404	679,677	376,989
Number of Schools	13,272	5,719	10,703	5,285	12,375	5,840

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.

Table A14: **Estimated Effects of Instructional Time by Quartile of Classroom Quality in Mathematics, Science, and Language Arts (2012-2018)**

	2012		2015		2018	
	(1)	(2)	(1)	(2)	(1)	(2)
Panel (a): Weekly Hours of Instruction						
Instruction time	0.005	0.026**	0.017***	0.018*	0.006	0.005
	(0.004)	(0.011)	(0.004)	(0.011)	(0.004)	(0.010)
Instruction time × 2nd quality quartile	0.001	0.000	-0.001	0.015	0.010**	0.003
	(0.005)	(0.016)	(0.005)	(0.015)	(0.005)	(0.014)
Instruction time × 3rd quality quartile	0.008	0.004	0.004	0.018	0.007	0.001
	(0.005)	(0.017)	(0.005)	(0.016)	(0.005)	(0.016)
Instruction time × 4th quality quartile	0.006	0.013	-0.001	0.009	0.009*	0.020
	(0.005)	(0.016)	(0.005)	(0.016)	(0.005)	(0.016)
Panel (b): Weekly Number of Class Periods						
Instruction time	0.005	0.038***	0.015***	0.024**	0.006*	0.042***
	(0.004)	(0.012)	(0.003)	(0.009)	(0.003)	(0.011)
Instruction time × 2nd quality quartile	-0.000	0.005	-0.001	0.028**	0.009**	0.000
	(0.005)	(0.016)	(0.004)	(0.014)	(0.004)	(0.015)
Instruction time × 3rd quality quartile	0.007	0.002	0.002	0.029**	0.007	0.006
	(0.005)	(0.016)	(0.004)	(0.014)	(0.004)	(0.016)
Instruction time × 4th quality quartile	0.005	0.006	-0.003	0.020	0.006	0.017
	(0.005)	(0.016)	(0.004)	(0.014)	(0.004)	(0.016)
School-by-grade fixed effect	Y	N	Y	N	Y	N
School-by-subject fixed effect	N	Y	N	Y	N	Y
Sample Size	476,664	237,798	739,401	424,404	679,677	376,989
Number of Schools	13,272	5,719	10,703	5,285	12,375	5,840

Notes. Robust standard errors clustered at the school-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. **Data source:** The OECD and PISA.