INTERNATIONAL MONETARY FUND

Class Discipline, Class Size and Scholastic Achievement Across Countries

A Theoretical and Empirical View of Educational Production

Noam Gruber

WP/26/105

2026
MAY

WORKING PAPER

 WP/26/105

IMF Working Paper
Institute of Capacity Development

Class Discipline, Class Size and Scholastic Achievement Across Countries: A Theoretical and Empirical View of Educational Production

Prepared by Noam Gruber*

Authorized for distribution by Paul Cashin
May 2026

ABSTRACT: Formulating an education production function and using estimates of student and class discipline levels, this paper seeks to identify the relations between discipline, class size, teaching quality and scholastic outcomes. The data shows both individual and class-level discipline to be a powerful predictor of the PISA math score, while variance in discipline among classmates has a strongly negative effect. Furthermore, class discipline is correlated with larger classes. As a structural simulation demonstrates, the correlations observed in the data can be well explained by how schools allocate students and teachers to classes. This analysis allows for a break-down of the contribution of educational production factors and highlights the role of individual and class discipline to student achievements.

RECOMMENDED CITATION: Noam Gruber. 2026. “Class Discipline, Class Size and Scholastic Achievement Across Countries: A Theoretical and Empirical View of Educational Production.” IMF Working Paper 2026/105, International Monetary Fund, Washington, DC.

JEL Classification Numbers:	I21, I28
Keywords:	Education Production, PISA, International Tests, Discipline, Tardiness, Truancy, Class Size, Teaching Quality
Author’s E-Mail Address:	ngruber@imf.org

* Noam Gruber is Resident Advisor of the IMF – Singapore Regional Training Institute.

WORKING PAPERS

Class Discipline, Class Size and Scholastic Achievement Across Countries

A Theoretical and Empirical View of Educational Production

Prepared by Noam Gruber[1]

[1] I am grateful to Dan Ben-David, Ayal Kimhi and Ity Shurtz and Rachel Tan, as well as participants of the Israeli Economic Association's convention, for their helpful comments on an early version of the paper. I thank Francesco Luna and Stephanie Medina Cas of the IMF's Institute for Capacity Development for their useful comments. I have no relevant or material financial interests that relate to the research described in this paper. The views expressed in this paper are mine alone and do not represent the views of the IMF, its Executive Board, or IMF management. All errors are my own.

1 Introduction

Much like economic value creation, the educational process can also be formulated as a function, where some of the central factors of educational production could arguably be individual student ability, effort, peer effect and teaching quality. However, unlike factors of industrial production, the definitions and measurements of such educational production factors are not yet established and standardized.

This paper attempts to measure and use discipline as a factor of educational production, both at the individual level, where it is a proxy for student effort, and at the class level, where it reflects class study environment and peer effect, and explore its role in the context of educational production. The role of discipline in explaining the differences in scholastic achievement levels between educational systems has thus far been under-explored, mostly due to lack of a standard measure of discipline. Using student reporting on class atmosphere as well as truancy and tardiness data from the PISA (the OECD's Programme for International Student Assessment), this paper constructs measures of student- and class-level discipline. These measures are shown to strongly correlate with both class sizes and student PISA math scores.

Theoretical analysis suggests that these correlations stem from several factors: first, direct effects – the positive effects of individual effort, proxied by student-level discipline, and peer-effect, proxied by class-level discipline. In addition, a selection process through which higher-discipline students are more likely to be assigned to larger classes can create both a positive peer effect and a negative class-size effect (more students, more interruptions). Furthermore, the possible allocation of teaching resources (e.g. better teachers) to larger classes can create an additional layer of positive correlation between discipline, class size and student achievement. Finally, by enabling larger classes, high student discipline enables school systems to hire fewer teachers and be more selective with regard to their quality.

Due to the difficulty of directly measuring teaching quality, its effect is particularly hard to estimate. Lacking a good proxy for teaching quality, its omission from regression estimations is likely to bias results for the variables with which it is correlated. To examine the impact of this issue on the empirical analysis, this paper generates synthetic data using a structural model of educational production, simulating selection of students and teachers into classes. Results are shown to qualitatively replicate correlations seen in the data. The omission of teaching quality biases upwards the effects of some discipline measures, but does not change their sign. It does however create a powerful correlation between class size and student scores, as seen in the data.

Section 2 overviews the relevant literature. Section 3 presents a theoretical model. Section 4 describes the PISA data used in this paper. Section 5 focuses on the estimation of discipline, comparing classroom disciplinary atmosphere and student-level discipline, based on students' truancy and tardiness. Section 6 contains econometric analysis of educational production. Section 7 presents a simulation using synthetic data generated by a theoretical model and compares it to PISA data. Section 8 concludes.

2 Literature Review

Previous papers have used data from international surveys such as PISA in attempts to identify key elements of the 'education production function' and explain the variance in outcomes across students, schools,

education systems and countries, using student, family and school attributes (for a comprehensive review see Hanushek and Woessmann (2011)). This strand of literature has generally not used indicators of discipline levels. It is often the case in this literature that when differences across countries cannot be explained away with hard data, unmeasurable cultural differences are called upon (e.g. see Perera and Asadullah (2019) regarding Malaysia and Asadullah et al. (2020) regarding Vietnam).

While school discipline and its effect on student outcomes has long been explored (for a review of the topic, see Arum and Velez (2012)), it has mostly been studied separately from other factors influencing the educational process. In essence, measurement issues have hampered a more thorough and systematic empirical analysis of the effect of school discipline across classes, schools and countries, with respect to pertinent educational factors such as class size and teaching quality. In a relevant attempt, Zamarro et al. (2019) construct measures of student effort in answering the PISA test to explain score variation across countries. In Asadullah et al. (2021), student effort is found to play a significant role in explaining score variation among students in Bangladesh.

In conjunction, the 'class size' literature attempts to explain the apparent paradox of positive correlation between class size and student performance. This vast empirical literature seeks to deal with selection issues and properly identify the effect of class size on the quality of education. Some meta-analysis papers, such as Krueger (2003) find that reducing class size has a positive impact on student performance. Others, such as Hoxby (2000) and Hanushek (2003), find little to no effect. Of two exceptional papers in this literature, Angrist and Lavy (1999), which uses the maximum class size regulation in Israel for identification, and Krueger (1999), which uses random assignment, both find that reducing class size does have a small positive effect on scholastic achievements.

Lazear (2001) attempts to explain the positive correlation between class size and student performance using a theoretical model, arguing that for higher levels of discipline, larger classes may indeed be optimal and lead to better results. Lazear (2001) is seminal in presenting a model which ties student outcomes to class discipline, class size and teaching quality together in one theoretical framework, exploring the potentially complex co-relations between these factors of education production.

It could be argued that the vast majority of research on the effect of class size is done in a partial equilibrium framework. This literature focuses on overcoming the selection bias causing better students to be assigned to larger classes, and often assumes that changing class sizes has no impact on the overall quality of teaching.[1] Lazear (2001) breaks this mold by modeling the trade-offs between class sizes and teaching quality (i.e. selectiveness in hiring teachers), and how they depend on class discipline.

As both classroom discipline and teaching quality are hard to measure quantitatively (on measuring teaching quality, see Hanushek and Rivkin (2006)), empirical work on their co-relations is scarce. Jepsen and Rivkin (2009) argues that higher discipline enables larger classes, thus reducing the number of teachers needed, vacating resources for teaching quality and allowing schools to be more selective vis-à-vis the teachers employed. It could also be argued that higher student discipline would attract a higher caliber of teachers to specific schools or to the teaching profession in general.[2] The causal relationship between country- and class-level discipline and teacher quality are thus potentially complex and bi-directional.

[1] There are some exceptions, such as Jepsen and Rivkin (2009).

[2] For example, Hanushek et al. (2004) shows that student characteristics are important factors in the decisions of teachers whether to leave specific schools.

3 The Education Production Function

Equation 1 presents a student-level education production function:

$$E_{i,j} = Q_j^{\beta_0} P_i^{\beta_1} P_j^{\beta_2} X_i^{\gamma_x} Z_j^{\gamma_z} \tag{1}$$

where $E_{i,j}$ is the educational product of student i in class j (proxied by the PISA test score),[3] Q_j is the quality of teaching at class j, P_i is student i's level of discipline ($0 \leq P \leq 1$), and P_j is the classroom level of discipline, which is the product of the level of discipline of students in class j (including student i, for simplicity) as seen in Equation 2 below. X_i is the set of student i's educational attributes other than discipline (e.g. talent, motivation and parental assistance), and Z_j is the set of such attributes of the other students in class j.

As stated above, classroom discipline is modeled as a product of student discipline levels:

$$P_j = \prod_{i=1}^{N_j} P_i \tag{2}$$

where N_j is the number of students in class j.

The multiplicative form of P_j is inspired by Lazear (2001), which views $\bar{P}$ as the probability of the representative student not interrupting his/hers classmates, and $\bar{P}^N$ therefore as the probability of uninterrupted study in a classroom with N students. Importantly, Lazear (2001) considers the impact of average class discipline ($\bar{P}$) in conjunction with the number of student in the class (N), but does not take into account the variation in classmates' discipline levels (i.e. σ_P), which, as the multiplicative form of Equation 2 suggests, is of great consequence.

Q_j, teaching quality in class j, is defined here as orthogonal to P_j, class discipline. In other words, Q is teaching quality/effectiveness beyond inspiring/enforcing discipline.

As discussed in Section 1, inability to proxy for teaching quality may bias the estimates of other variables which due to potential selection may be correlated with it. Specifically, class size, class discipline and teaching quality may all be mutually correlated, as Section 7 demonstrates by simulation.

The mechanisms behinds such correlations are of great importance. Let us assume that school management seeks to maximize educational outcomes using the following two-step optimization procedure:

Firstly, it seeks to maintain classroom discipline levels above some implicit national/cultural threshold, so that $P_j \geq P_C$, where P_C is the national implicit classroom discipline threshold. Then, much along the lines of Lazear (2001)'s argument, class size (the number of students in the class) will be positively correlated with average class discipline ($N_j \propto \bar{P}_j$), where $\bar{P}_j$ is the simple average of student discipline in the class ($\bar{P}_j = \frac{\sum_{i=1}^{N_j} P_j}{N_j} \neq P_j$).[4]

Secondly, the school has access to discrete teaching resources (i.e. teachers) drawn from a given distribution.[5] After having allocated the more disciplined students to larger classes to meet the implicit classroom

[3] This is a vast simplification, as the educational product is a stock rather than a flow, and is clearly more complex and multi-dimensional than a single score.

[4] Notice that class size will not be correlated with P_j within a country (or cultural region), as all classes will have classroom discipline close to P_C.

[5] The simulation in Section 7 uses a Normal distribution: $Q \sim N(\bar{Q}, \sigma_Q)$). Notice the underlying simplifying assumption that all countries draw from a similar distribution of teacher quality.

discipline threshold P_C, school management also assigns higher quality teachers to larger classes. Teaching quality Q_j will then be positively correlated to class size N_j, average student discipline in the class ($\bar{P}_j$, which, again, is different from P_j) and scholastic outcomes in the class, i.e. $E_{i\in j} \propto Q_j \propto N_j \propto \bar{P}_j$.

As the empirical literature has rarely been able to measure teaching quality and/or class discipline, it is often found that class size is counter-intuitively correlated with students' performance. By empirically estimating discipline, this paper can look at the role played by schools sorting students to classes according to discipline levels in creating this correlation. Given perfect information, we would have liked to estimate the following model, based on Equation 1 in log form:

$$e_{i,j} = \alpha + \beta_0 q_j + \beta_1 p_i + \beta_2 p_j + \gamma_x x_i + \gamma_z z_j + \varepsilon_{i,j} \tag{3}$$

Note that p_j is the log form of P_j, hence $p_j = \sum_{i=1}^{n_j} p_i = (n_j)\bar{p}_j$.

4 PISA Data: 2012, 2015, 2018 and 2022 Rounds

The PISA test is administered to 15 year olds[6] every three years (the round planned for 2021 was postponed to 2022 due to the COVID-19 pandemic).[7] This paper uses data from the 2012, 2015, 2018 and 2022 rounds. Of the three main subjects covered by PISA –- mathematics, literacy and science –- math was chosen as the focus. Beyond its intrinsic importance, math is an international language and is therefore uniquely suited for cross-country comparison.[8] It should be noted that there is a high correlation between countries' math achievements and their attainments in literacy and science, meaning that an education system's success in teaching quantitative thinking skills is a good indicator of its success in the teaching of other skills.

The PISA exams were calibrated in 2000 to an OECD country average of 500 and standard deviation of 100. All subsequent rounds were calibrated with respect to the 2000 round using Item Response Theory (IRT) methodology. In addition to the test questions, PISA respondents, as well as school administrators, also fill out survey questionnaires, which provide many details on the students' family background, school attributes, attitudes etc. The present study uses data for the countries considered advanced economies by the IMF at 2012, the first survey year, minus some exceptionally small countries (Cyprus, Iceland, Luxembourg, Malta and San Marino), 30 countries in total.[9] For these 30 countries the data comprises 247,985 test-takers in 2012, 223,611 test-takers in 2015, 253,680 test-takers in 2018 and 251,385 test-takes in 2022.

In addition to PISA data, this paper uses World Bank data on GDP per capita and the share of population 14 and under, with CEIC data used for Taiwan. These variables are used as controls to supplement PISA data.

[6] The cohort of 15 years and 3 months old to 16 years and 2 months old.
[7] See the OECD website: http://www.oecd.org/pisa/aboutpisa.
[8] For sample questions from the PISA exam, see: http://www.oecd.org/pisa/test.
[9] For the list of advanced economies see https://www.imf.org/external/pubs/ft/weo/2012/02/weodata/groups.htm.

5 Class Level and Student Level Discipline Measures

5.1 Classroom Disciplinary Atmosphere

In the PISA questionnaire three statements were posed to the students that relate directly to classroom discipline levels: "Students don't listen to what the teacher says," "There is noise and disorder," and "The teacher has to wait a long time for students to quiet down." For each of these statements, the students had to mark one of the following responses: "Every lesson," "Most lessons," "Some lessons," and "Never or hardly ever."

Figure 1 demonstrates the correlation between the perceived classroom disciplinary climate reflected by the response to these statements and the students' average score, compared to the respective national average. This correlation is clear, with students who perceive a good disciplinary climate achieving higher average scores.

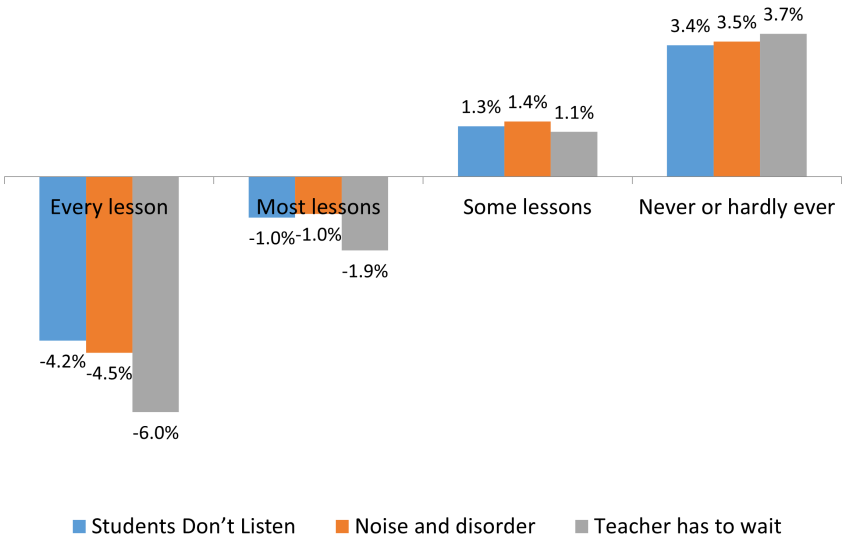


Figure 1: Percent difference from country average math score, by classroom discipline measure

On the assumption that cultural bias is relatively small within a given education system (in contrast to bias between different education systems, which can be considerable), it is not surprising that these figures suggest a major advantage for students in classrooms where perceived discipline levels are high.

Using the factor analysis method, the responses to the three statements above ("Students don't listen to what the teacher says," "There is noise and disorder," and "The teacher has to wait a long time for students to quiet down") are reduced to a single 'classroom discipline' index, which is normalized to the 0.01-1 range.[10] Country ranking and index scores are presented in Figure A1 in the appendix.

[10]The 0.01-1 range is used in reference to the concept of discipline as the probability of non-interruption in the classroom, as discussed in Section 3. However, while the index can arguably be viewed as a monotonous transformation of the probability of non-interruption, it is not identical to it.

There are two important points to make regarding this 'classroom discipline' index: the first, that it attests to the students' perception of class atmosphere, which is a product of overall student behavior in the class, and not to their own individual behavior. In other words, it contains information on P_j but, does not relate directly to P_i or $\bar{P}_j$. The second point is that it may be subject to cultural bias, as it is influenced by students' perception of discipline in each locale (later in the analysis this potential cultural bias is partially accounted for using country fixed effects).

Figure 2 displays the 'classroom discipline' index versus PISA math score (both averaged over the 2012, 2015, 2018 and 2022 rounds), by country.

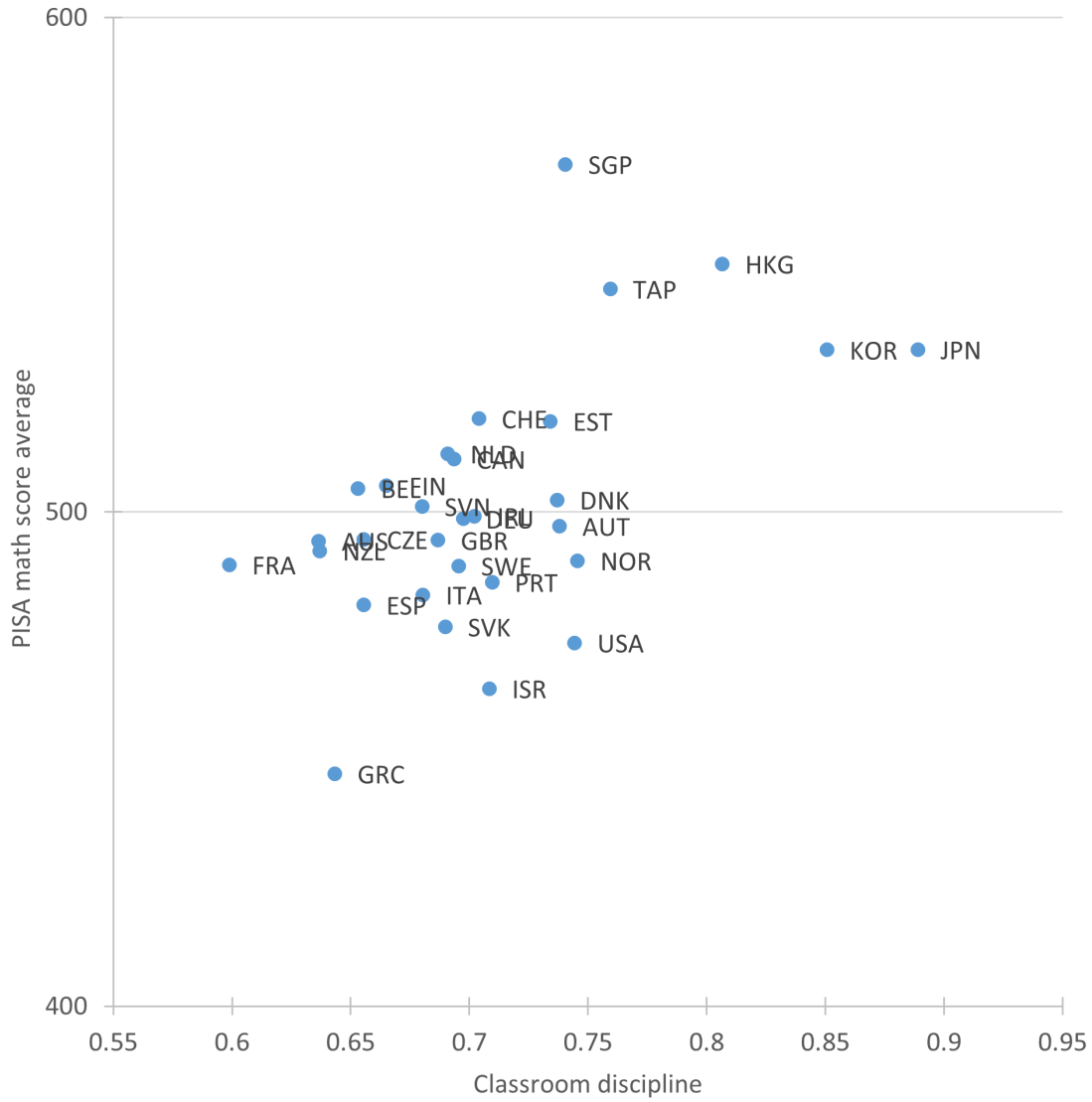


Figure 2: PISA score versus classroom discipline, by country (2012, 2015, 2018 and 2022 test rounds average)

Of the 5 leading Asian countries, Hong-Kong, Korea and Japan enjoy high scores in conjunction with high classroom discipline, while Singapore and Taiwan have very high scores but only moderately-high levels of classroom discipline. Of the low scoring countries, Greece, and to a lesser degree Slovakia and Spain, suffer from both low scores and low classroom discipline. However, the correlation is far from perfect. Norwegian students report a high level of classroom discipline, yet their test scores are comparable to those of their French counterparts, who report the lowest levels of classroom discipline in the developed world. Israeli and American students report a low-moderate and high-moderate classroom disciplinary climate respectively, while their average test scores are fairly low. Belgian students report a weak classroom disciplinary atmosphere while scoring above average.

5.2 Individual Discipline: Student Discipline as Reflected by Truancy and Tardiness

Unlike the classroom discipline measure, the individual discipline measure is based on self-reported truancy and tardiness data. Truancy and tardiness have a direct negative impact on the students themselves, as they lose study time; they also affect the entire class by creating gaps in knowledge that have to be addressed. But beyond this, data on truancy and tardiness also testify to individual and school motivation and discipline levels.[11] The PISA questionnaire asks students about the number of times they arrived late to school, the number of their full-day absences, and the number of times they skipped class (without skipping the entire school day), during the two weeks preceding the test.

As with the statements regarding classroom discipline, the data on truancy and tardiness strongly correlate with student exam scores. This can be seen in Figure 3.

Using the factor analysis method, the tardiness and truancy data (late arrivals, full-day absences and skipping classes) are reduced to a single 'individual discipline' index, which, similarly to the classroom discipline index, is normalized to the 0.01-1 range. Country ranking and index scores are presented in Figure A2 in the appendix.

Figure 4 displays the individual discipline index versus PISA math score (both averaged between the 2012, 2015, 2018 and 2022 rounds), by country. A partial correlation can be readily detected between individual discipline and country score. Of the top performing countries, the 5 Asian nations, Switzerland and the Netherlands all exhibit very high individual discipline levels; by contrast, individual discipline levels in Estonia and Canada are low to average, despite their relatively high scores. Some of the low-scoring countries, such as Italy, Israel and Greece, and to a lesser degree Spain, Portugal, Slovakia and the US, are characterized by low individual discipline levels.

The individual discipline index reinforces many cultural stereotypes: the East Asian and Northern European countries are at the top of the ladder, while the Mediterranean countries (and Portugal) occupy the bottom rungs. It is important to note that this index contains information on the individual student's level of discipline (P_i) but, as can be seen in Equation 2, its influence on class-level discipline (P_j) is multiplicative and dependent on the number of students in the class.

[11] The effects of truancy on scholastic achievement have been explored in papers such as Bosworth (1994) and Goodman (2014).

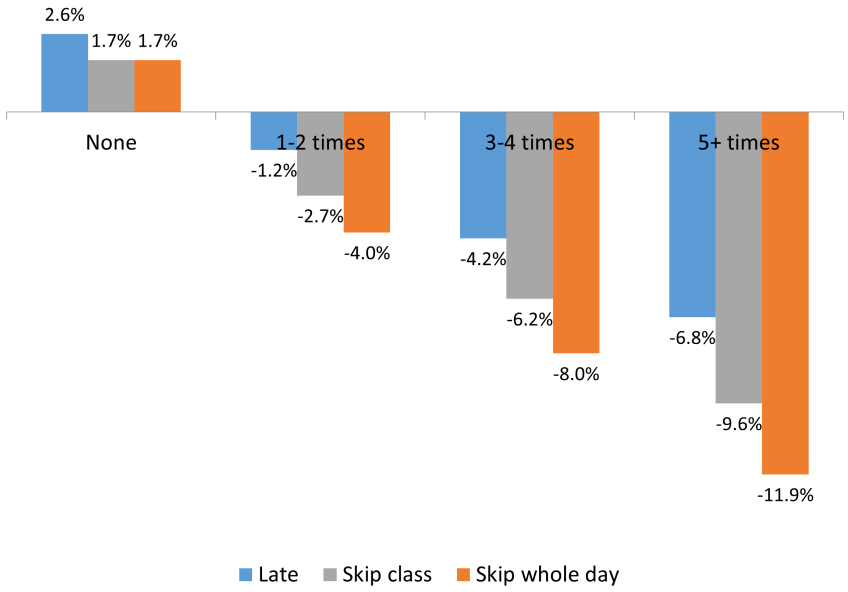


Figure 3: Percent difference from country average math score, by individual discipline measures (truancy and tardiness during the last two weeks)

5.3 The Relation Between Individual and Classroom Measures of Discipline

As Figure 5 demonstrates, the picture that emerges from truancy and tardiness data (reflected by the individual discipline index) can be different from the one painted by students' impressions regarding classroom discipline. While some countries are located close to the diagonal of the graph, meaning that their classroom and individual indices are similar, for some countries this is clearly not the case. Italian and Israeli students report a disciplinary climate much better than what the truancy and tardiness data indicate, while for France, Belgium and the Czech Republic the opposite is true.

As discussed in Section 3, the effect of individual discipline on classroom discipline is multiplicative. Equation 2 postulates that $P_j = \prod_{i=1}^{N_j} P_i$. According to this, the relation between the individual discipline index (which correlates to P_i) and the classroom discipline index (which correlates to P_j) is influenced both by N_j, the class size, and the variation in the levels of individual student discipline within the class. I.e., a class with the same average student discipline ($\bar{P}_j = \frac{\sum_{i=1}^{N_j} P_i}{N_j}$) will have a lower discipline atmosphere (P_j) if the number of students is larger (creating more interruptions) or if the within-class variation in student discipline (σ_{P_j}, the standard deviation of P in class j) is greater (due to the disproportional effect of students with low discipline, as represented by the multiplicative form of Equation 2). Finally, cultural bias in the perception of discipline can also account for some of the gaps between classroom discipline (P_j, based on student reporting on disciplinary atmosphere) and average individual discipline ($\bar{P}_j$, based on student truancy and tardiness self-reporting).

Table 1 examines the relation between classroom discipline (P_j), average individual discipline ($\bar{P}_j$) and the standard deviation of individual discipline (σ_{P_j}), at the school level. The inclusion of country fixed effects accounts for any country-level cultural bias in the perception of discipline. Another obfuscating factor is the

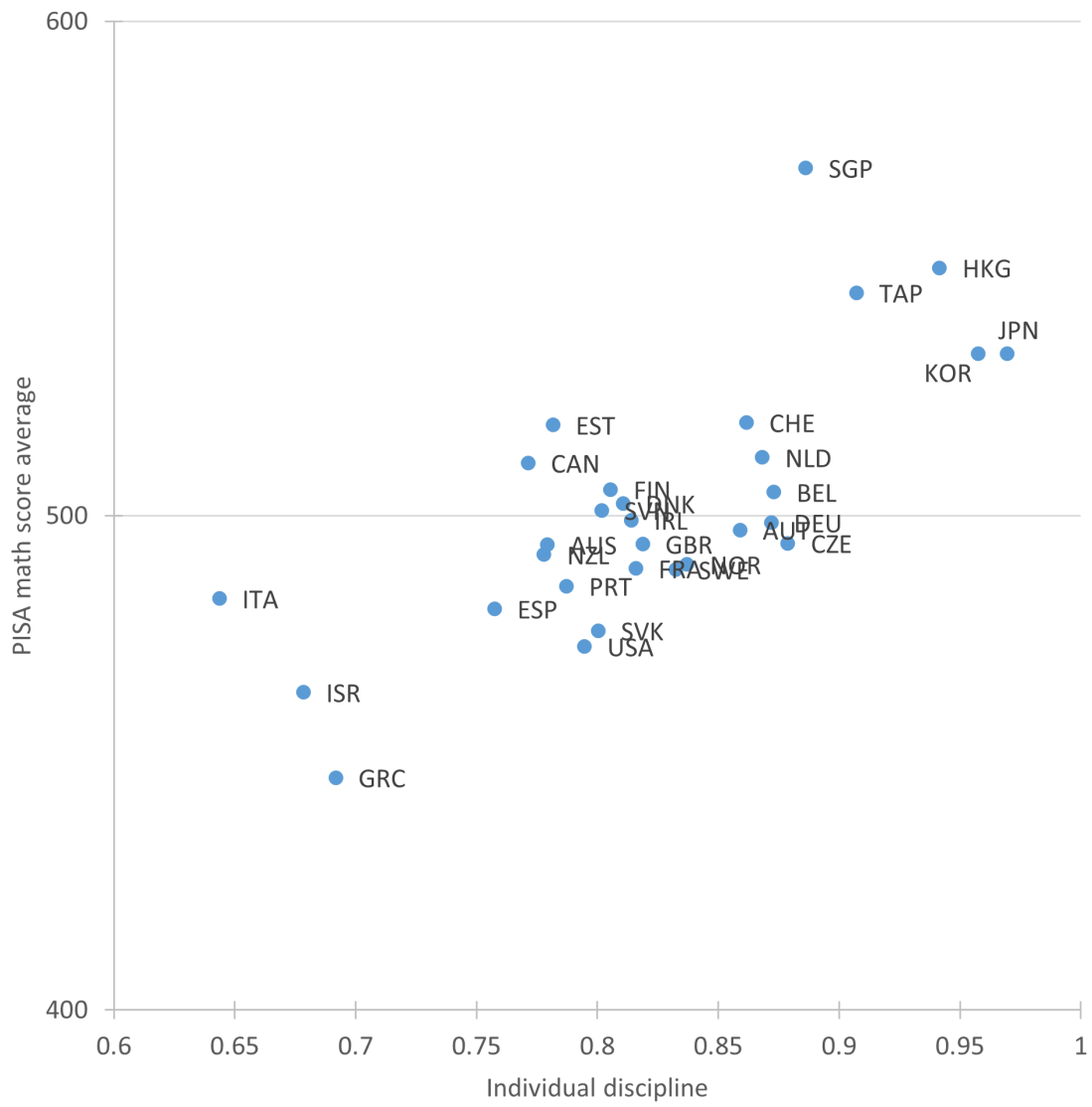


Figure 4: PISA score versus individual discipline (2012, 2015, 2018 and 2022 test rounds average)

postulated bi-directional relation between individual discipline and class size: on the one hand, class size amplifies the effect of average individual discipline on classroom discipline; on the other hand, class size is endogenous to students' discipline (as discussed in detail in Subsection 5.4).

As can be seen in Column (1), the average individual discipline index coefficient is positive and statistically significant at the 1% level (a t-value of 6.62). In addition, the (school-level) standard deviation is negative and statistically significant at the 1% level (a t-value of -6.38). The R^2 is 0.202. The positive coefficient for the average and negative coefficient for the standard deviation are in line with the Equation 2 formulation and its underlying logic: one troublesome student can bring down an entire class.

To deal with the issue of class size endogeneity, Columns (2)-(7) estimate the same specification for the

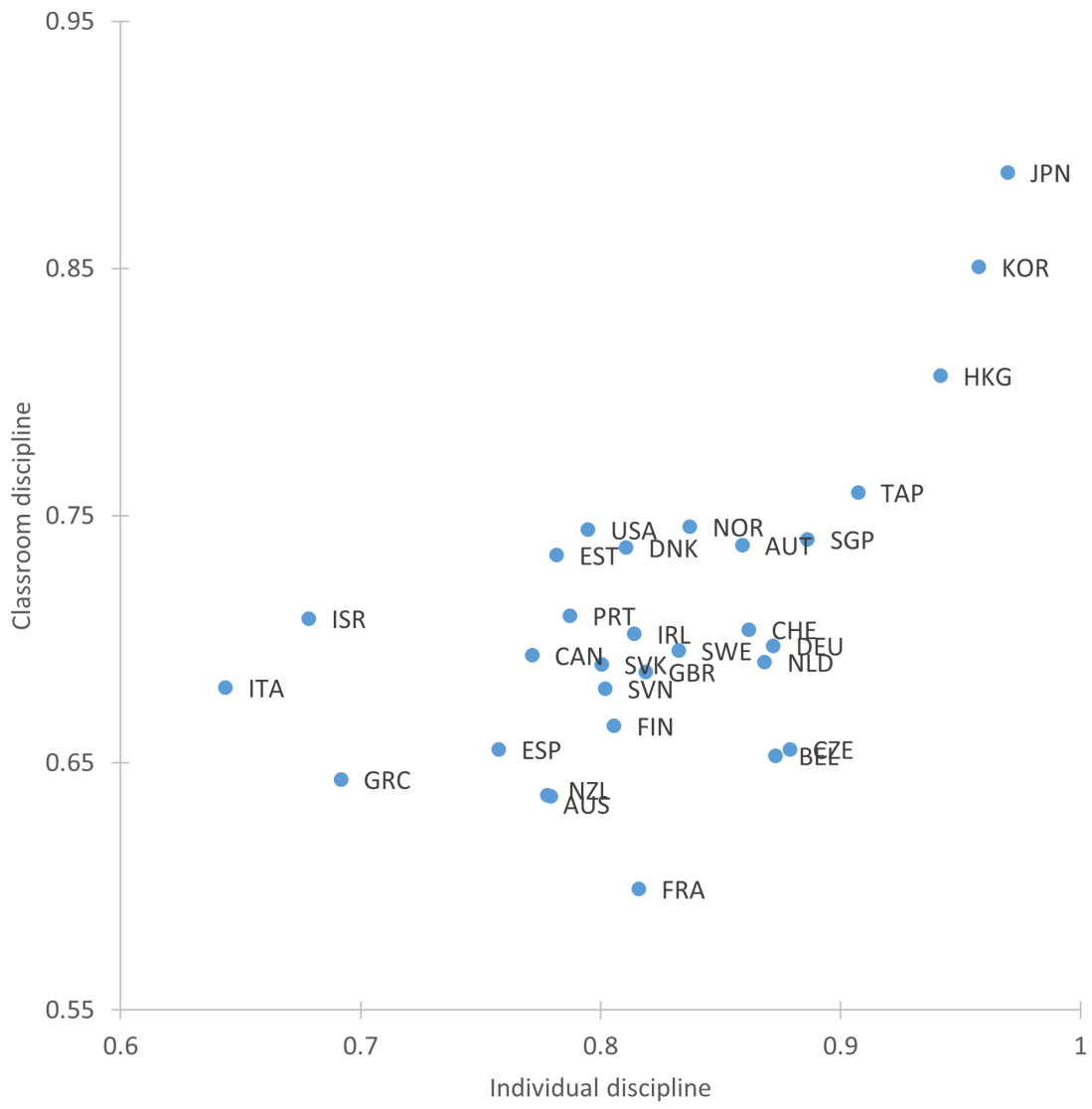


Figure 5: Classroom versus individual discipline, by country (2012, 2015, 2018 and 2022 test rounds average)

most common class sizes separately (for median class sizes by country and year, see Table A1). Results are qualitatively quite similar: in all regressions the coefficients of the average and standard deviation of individual discipline maintain their positive and negatives signs respectively. The coefficients for the standard deviation of individual discipline in Column (2) and (3) (class sizes -15 and 16-20) are statistically significant only at the 10% and 5% level, respectively. All other coefficients in Columns (2)-(7) are significant at the 1% level. All regressions control for year and country fixed effects, as well as GDP per capita and share of population under 15.

Table 1: Classroom discipline (P_j) versus individual discipline ($\bar{P}_j$ and σ_{P_j}), school-level analysis

	All Class Sizes	15 or less	16-20	21-25	26-30	31-35	36-40
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\bar{P}_j$ (mean ind. disc.)	.260 (.039)***	.263 (.057)***	.352 (.039)***	.274 (.030)***	.252 (.048)***	.236 (.105)**	.429 (.063)***
σ_{P_j} (SD ind. disc.)	-.193 (.030)***	-.125 (.059)**	-.076 (.040)*	-.233 (.037)***	-.175 (.043)***	-.317 (.068)***	-.224 (.082)***
Const.	.704 (.098)***	.733 (.197)***	.559 (.127)***	.562 (.121)***	.643 (.133)***	1.133 (.215)***	1.056 (.229)***
Obs.	35796	1819	5311	11002	8566	2083	1335
R^2	.202	.131	.164	.186	.212	.39	.525

Notes: All regressions control for year, country, country GDP per capita and country share of population under age 15. Errors clustered by country.
*Significant at 10%; **significant at 5%; ***significant at 1%

Clearly, students' truancy and tardiness data provide strong indications for classrooms' disciplinary climates.

5.4 Discipline and Class Size

The theoretical model presented in Section 3 predicts class size to have a magnifying effect on the relation between average individual discipline ($\bar{P}_j$) and student achievements (see Equation 2). In that sense, for a given level of individual discipline and teaching quality, we would expect an increase in class size to lower test results. However, the model also assumes that school management adjusts student number and allocation in order to pass a class discipline threshold, and then allocates more resources (i.e. better teachers) to the larger classes. Such optimizations would yield a positive correlation between individual discipline (P_i) and class size, and would extend to teaching quality and educational outcomes. Importantly, classroom discipline (P_j) would not be a correlated to class size, as all classes in a given education system (country/region/stream) are constructed to meet the same classroom discipline threshold.

Figure 6 presents individual discipline level (P_i) versus class size by country averages. As can be seen, the Asian 5 leading countries display high discipline levels and large classes, while most Western countries are characterized by smaller classes and differing levels of discipline. Of the latter, countries such as Austria, Belgium, the Czech Republic, Germany, the Netherlands and Switzerland enjoy relatively high levels of student discipline. Arguably, it could be more optimal for them to have larger classes, which would allow for increasing teaching quality through greater selectivity in hiring.

Of the low-performing countries, Israel has fairly large classes and a low-level of individual discipline, clearly a bad combination. Greece and Italy have medium sized classes but low discipline levels. It can

therefore be argued that in order to improve their achievements, the school systems in these countries need to either lower class sizes, at a considerable cost (financially, and in terms of teachers' quality), or increase discipline. The ability of educational systems with large classes and low levels of discipline to attract and maintain high quality teaching staff is also of great concern (not discussed in this paper).

It is clear from Figure 6 that at the country-level class size is far from fully-determined by the level of country-average individual discipline. Factors such as budget constraints, demographic trends and cultural preferences undoubtedly contribute to this country-level variation in class size. Nevertheless, individual discipline is strongly correlated to within-country variation in class size, as Table 2 demonstrates.

Table 2 presents school-level regressions of class size versus classroom discipline (P_j) and individual discipline (school mean, $\bar{P}_j$, and standard deviation, σ_{P_j}), in all possible combinations. In all instances in which the respective variables are included, mean individual discipline ($\bar{P}_j$) is statistically significant at 1%, while classroom discipline (P_j) and the standard deviation of individual discipline (σ_{P_j}) are not statistically significant even at 10%. This reflects the strong correlation between mean individual discipline ($\bar{P}_j$) and class size, as predicted by theory: a high level of average discipline enables effective teaching in large classes. The lack of correlation of class size with classroom discipline (P_j) is also in line with the theoretical model: if all classes are set up to meet the same classroom discipline threshold, there will be no correlation between classroom discipline and class size (whatever country differences in the threshold will be captured by country fixed effects).

The fact that the standard deviation of individual discipline (σ_{P_j}) in not correlated to class size is somewhat surprising: based on the model, one would assume high variability of students' discipline levels will induce schools to reduce class size. A possible explanation is that schools tolerate low discipline up to a certain level, above which they would move the student to a smaller, more suitable class.

Figure 7 presents class sizes and PISA math scores across countries (2012, 2015, 2018 and 2022 averages). Clearly, class size by itself is not a sufficient indicator of PISA scores at the country level. For example, Israel, France, Korea and Hong-Kong all have fairly similar class sizes, yet their PISA scores differ greatly. As shown later, these differences can at least partly be explained by the different levels of discipline in these countries. For example, the 5 Asian countries combine large classes with high student discipline, and are thus able to maintain good classroom discipline while arguably being more selective in teacher hiring, leading to high scores. On the opposite side of the spectrum we find countries such as Israel and Greece, which have large classes despite low (Israel) to medium (Greece) levels of student discipline, and

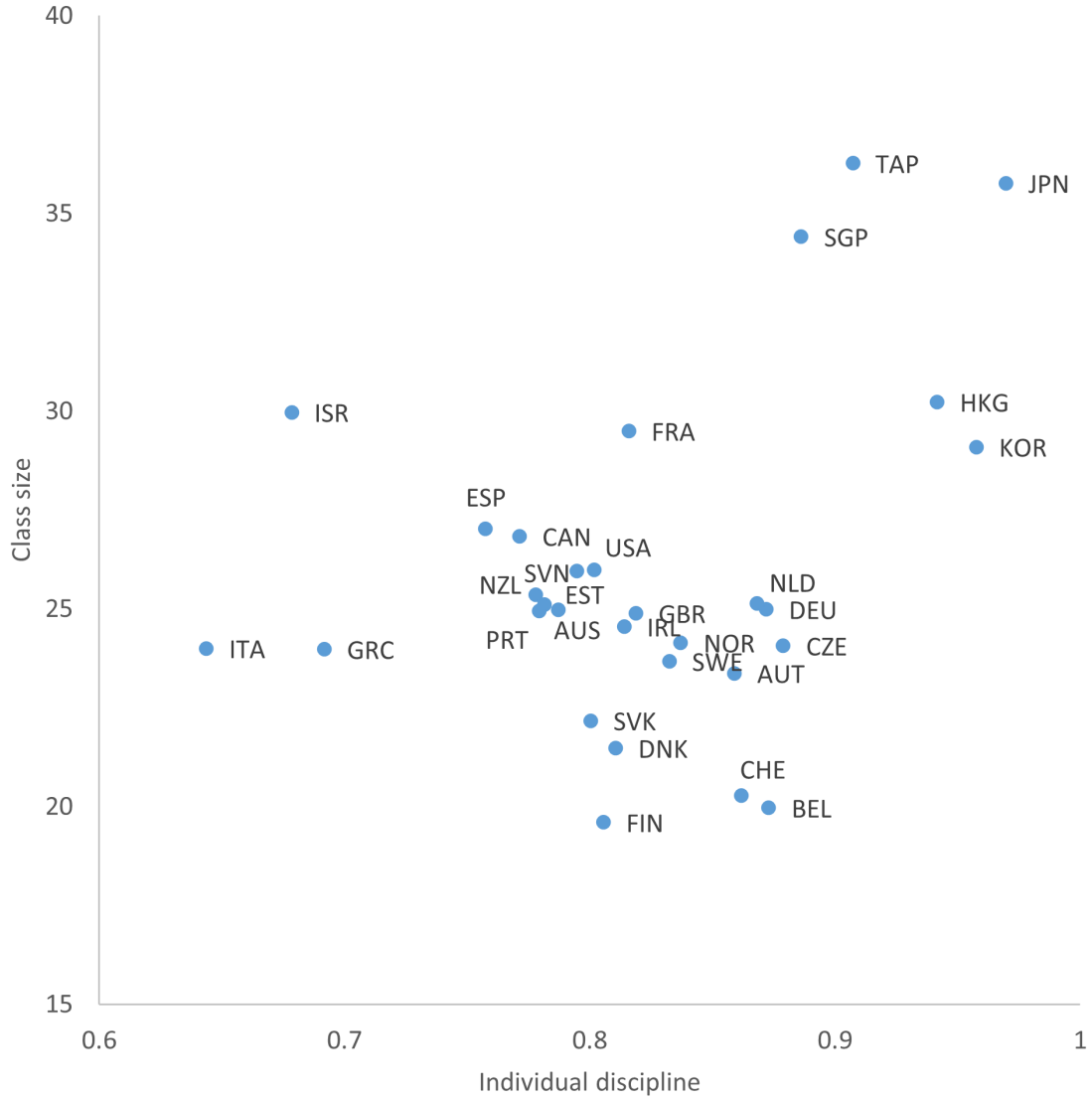


Figure 6: Individual discipline (P_i) versus class size, by country (2012, 2015, 2018 and 2022 test rounds average)

Table 2: Class size versus classroom discipline (P_j) and individual discipline ($\bar{P}_j$ and σ_{P_j}), school-level analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
P_j (class. disc.)	-.260			-.904	-.564		-.863
	(.654)			(.598)	(.650)		(.616)
$\bar{P}_j$ (mean ind. disc.)		3.042		2.973		4.519	4.237
		(1.140)***		(1.122)***		(1.366)***	(1.268)***
σ_{P_j} (SD ind. disc.)			-2.068		-2.118	2.192	1.762
			(1.472)		(1.519)	(1.438)	(1.507)
Const.	15.373	12.559	15.394	13.715	16.304	10.924	12.379
	(8.616)*	(9.065)	(8.674)*	(9.030)	(8.651)*	(8.983)	(8.918)
Obs.	31386	31576	31313	31147	30948	31313	30948
R^2	.276	.276	.275	.276	.275	.277	.277

Notes: Controls for year, country, country GDP per capita and country share of population under age 15. Errors clustered by country.
*Significant at 10%; **significant at 5%; ***significant at 1%

are clearly paying the price in the form of low educational attainments.

6 Estimating Education Production

This section is attempting to estimate the education production model at the student level, while acknowledging the potential bias due to the omission of teaching quality.

The model presented in Section 3 predicts individual discipline to be correlated with class size, and the two to be correlated with unobserved teaching quality, and therefore with student outcomes. This framework yields a positive relation between individual discipline, class size and student outcomes. The positive correlation between individual discipline and student outcomes can be attributed both to the direct effect (individual discipline to individual outcome) and to indirect effects (correlation with unobserved teaching quality). Notice that the positive relation between class size and student outcomes is predicted to be purely through correlation with latent teaching quality, as the negative effect of class size is offset by higher average individual discipline through student selection.

As a result of the these correlations, the omission of (unobserved) teaching quality is predicted to bias the effects of discipline (individual (P_i) and class average ($\bar{P}_j$)) and class size upwards, as higher discipline students are argued to be assigned to larger classes which in turn are assigned better teachers.

Teachers in PISA data are not allocated randomly, or according to any well defined rules, and there are no clear shocks to the allocation process to serve as natural experiments. To deal with this omitted variable bias, this paper takes two approaches: first, Section 7 gauges the bias by using simulation-based synthetic

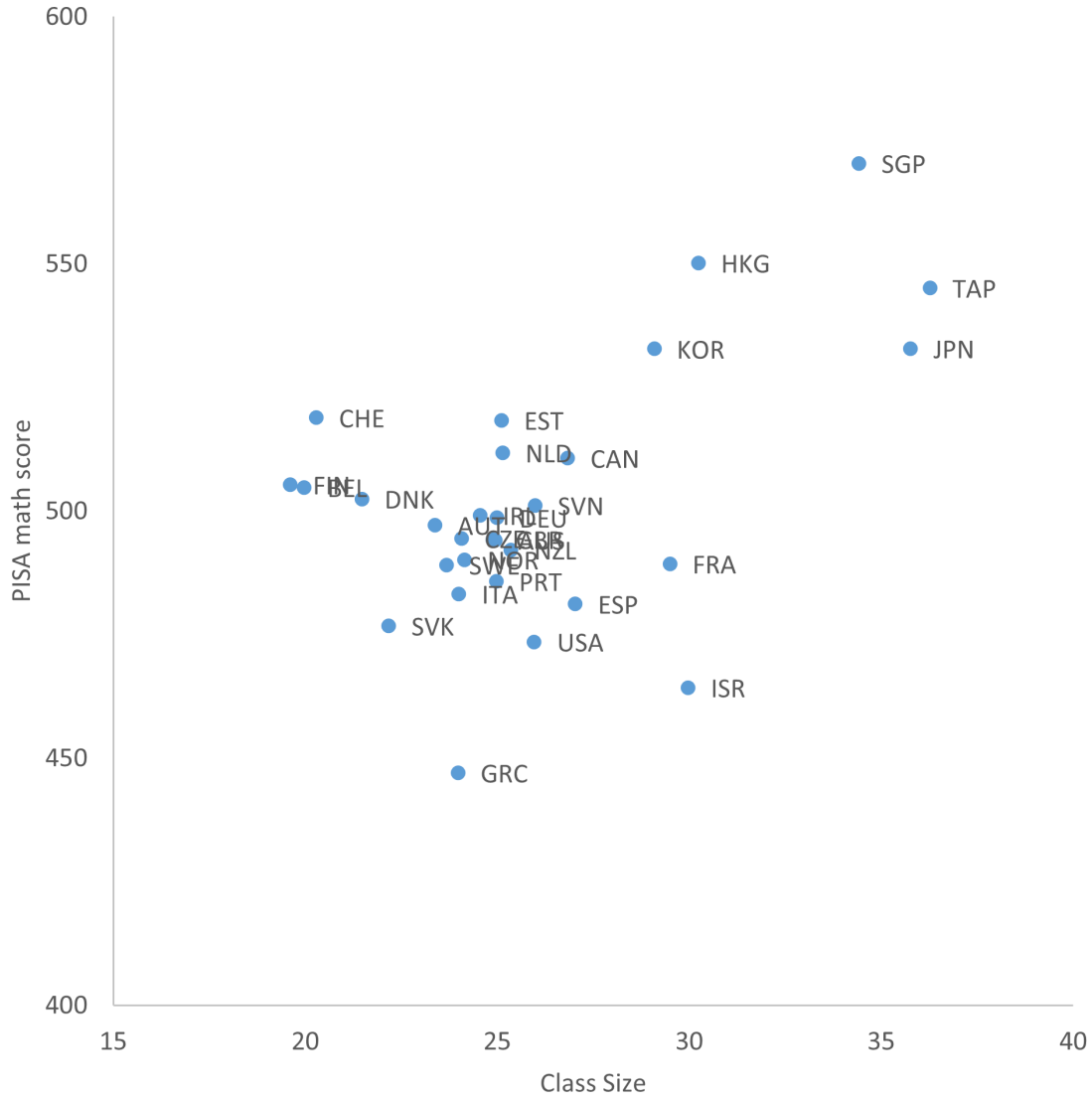


Figure 7: PISA math scores versus class size, by country (2012, 2015, 2018 and 2022 test rounds average)

data to compare estimations including and excluding teacher quality; second, general estimation results are compared to estimations using fixed class sizes.

Table 3 presents estimations of different specifications, correlating individual student PISA scores with class sizes, while incrementally introducing individual discipline P_i, average individual discipline (at the class level) $\bar{P}_j$, the standard deviation of individual discipline (at the class level) σ_{P_j} and class discipline P_j. All specifications control for year, country, country GDP per capita, country share of population under age 15, maternal and paternal education and share of schoolmates whose parents have post-secondary education.

Column (1) serves as a baseline, adding only class size fixed effects dummies. These dummies are all statistically significant at 1%, and indicate that an increase in class size is correlated with increase in student PISA scores up until a class size of 36-40, after which the positive effect (versus a baseline class of up to 15 students) starts to diminish. An intuitive explanation is that classes with more than 40 students are not normative for many school systems, and thus may indicate a school in financial and/or staffing trouble. In addition, as discussed in Section 7, the process of schools assigning students to classes using a noisy measure of student individual discipline may bias the school estimate of P_j upwards, leading to large classes where the students' high level of individual discipline does not fully offset the negative effect of class size on classroom discipline.

Column (2) adds individual discipline (P_i). The student's level of discipline is strongly correlated with his/her PISA score, with a coefficient of 67.25 and a t-value of 21.84. With the inclusion of individual discipline, the coefficients of the class size fixed effect dummies all drop in size, though they remain statistically significant at 1%. This makes sense, as individual discipline is correlated with class size, as seen in Table 2. Compared to Column (1) the R^2 improves from 0.227 to 0.255, indicating the substantial explanatory value of individual discipline.

Column (3) adds class average discipline ($\bar{P}_j$) to the regression, which is shown to be strongly correlated to score, with a coefficient of 115.02 and a t-value of 5.03. The coefficient of individual discipline (P_i) drops to 54.31 but remains highly significant, with a t-value of 14.35. These results separate individual discipline from the peer effect, showing them both to be strong factors. R^2 increases to 0.264.

Column (4) adds σ_{P_j}, the standard deviation of class discipline (the standard deviation of individual discipline within the class). As can be expected, greater variance in student discipline in the class is negatively associated with student scores. For a given class size and average class discipline, greater variance implies more relatively low-discipline students, having an out-sized negative effect on class atmosphere, as highlighted by the multiplicative nature of class discipline (see Equation 2 in Section 3). This correlation is significant at the 5% level (t-value of -2.56). Note however that average class discipline drops in size and is significant only at the 10% level (t-value of 1.78).

Column (5) adds class discipline (P_j). The coefficient of P_j is large (71.494) and highly significant (a t-value of 12.73). The addition of P_j diminishes the coefficient $\bar{P}_j$ and causes it to lose statistical significance even at the 10% level. This makes sense since the two variables are correlated, as Table 1 shows. While the coefficient of σ_{P_j} is also reduced, it remains negative, sizable and statistically significant at the 5% level (t-value of -2.2). Notice also that R^2 increases to 0.274.

Table 3: PISA math scores vs. measures of discipline, student-level analysis

	(1)	(2)	(3)	(4)	(5)
P_i (ind. disc.)		67.252 (3.079)***	54.309 (3.783)***	54.345 (3.796)***	54.429 (3.802)***
$\bar{P}_j$ (class mean ind. disc.)			115.017 (22.854)***	68.158 (38.388)*	49.170 (36.579)
σ_{P_j} (class SD ind. disc.)				-91.462 (35.668)**	-77.079 (35.061)**
P_j (class discipline)					71.494 (5.616)***
Class size					
16-20	13.817 (3.527)***	8.922 (2.527)***	8.814 (2.222)***	9.139 (2.093)***	8.140 (1.982)***
21-25	20.064 (4.316)***	14.886 (3.326)***	14.968 (3.139)***	15.218 (3.006)***	14.617 (2.926)***
26-30	28.855 (5.600)***	22.909 (4.129)***	22.171 (3.805)***	22.388 (3.753)***	21.999 (3.365)***
31-35	38.853 (7.188)***	32.219 (5.717)***	31.279 (5.177)***	31.473 (5.161)***	31.387 (4.810)***
36-40	43.945 (7.072)***	38.236 (5.818)***	37.737 (5.583)***	37.491 (5.451)***	37.284 (5.084)***
41-45	42.150 (8.827)***	36.609 (7.796)***	35.788 (7.786)***	35.461 (7.761)***	34.082 (7.574)***
46-50	22.375 (7.827)***	17.698 (6.362)***	18.028 (6.003)***	18.400 (6.003)***	16.753 (5.211)***
51+	13.762 (4.010)***	8.912 (2.972)***	8.865 (3.137)***	9.019 (3.101)***	8.928 (2.497)***
Const.	372.828 (47.876)***	322.034 (45.479)***	235.757 (47.078)***	290.425 (60.629)***	246.722 (56.617)***
Obs.	782389	743639	743639	743546	740340
R^2	.227	.255	.264	.266	.274

Notes: Controls for year, country, country GDP per capita, country share of population under age 15, maternal and paternal education and share of schoolmates whose parents have post-secondary education. Errors clustered by country.
*Significant at 10%; **significant at 5%; ***significant at 1%

As discussed above, class size has an important and potentially endogenous effect on the relation between (individual and classroom) discipline and scholastic achievement. As a robustness check, Tables A2 and A3 in the appendix examine the effect of class size on this relation.

Table A2 compares general estimation results (Column (1) in Table A2 is identical to Column (5) in Table 3) to estimations using fixed class sizes. While results are quite similar, it is interesting to note that the size of the negative effect of σ_{P_j} (the standard deviation of individual discipline in the class) is generally larger and more statistically significant for larger classes (except for the 51+ size, which is a small outlier).

Table A3 looks at interactions between class size and different measures of discipline, and their correlation with PISA math scores. Column (1) presents interactions with P_i, individual discipline. Column (2) presents interactions with $\bar{P}_j$, school mean individual discipline. Column (3) presents interactions with σ_{P_j}, the standard deviation of individual discipline in the class. Column (4) presents interactions with P_j (classroom discipline). Of the different measures, the interaction terms of σ_{P_j} with class sizes (Column (3)) is the

most statistically significant. Here too the negative effect is larger for bigger classes.

Tables A2 and A3 indicate that the correlations between student scores and discipline measures are largely stable in different class sizes, except for σ_{P_j}, the standard deviation of individual discipline in the class, where the negative effect seems to be growing in class size. However, these robustness checks do not resolve the issue of teaching quality as a latent variable possibly correlated with both individual discipline and class size. Section 7 looks at this issue by generating synthetic data using the theoretical model discussed in Section 3.

7 Model and Simulation

7.1 Synthetic Data Generation

To investigate both whether the model-suggested data generating process can create the correlations we see in the data, and the potential bias caused my the omission of unobserved teaching quality, synthetic data is generated as follows:

For $N_{cnt} = 30$ countries, an average country discipline is drawn from a global Normal distribution: $D_{cnt} \sim N(D_{world} = 1.5e, \sigma_{world} = 0.15e)$. Each country has $N_{stu} = 500$ students (essentially one large school with multiple classes). Students draw individual discipline levels based on the country draw such that $D_{stu} \sim N(D_{cnt}, \sigma_{stu} = 0.2e)$. For simplicity, σ_{stu} (the standard deviation of student discipline) is uniform across countries.

School management has noisy information on the students' level of discipline, such that $D_{sch} \sim N(D_{stu}, \sigma_{sch} = 0.3e)$. For simplicity, σ_{sch}, the noise in the school measurement of student discipline, is also uniform across countries. Student discipline values (normally distributed) are mapped to probabilities of non-disruption using an inverse-logit transformation, so that $P_i = invlogit(D_i)$.[12]

Classes are formed in the following manner:

Class sizes (n_{class}) range from $CS_{min} = 10$ to $CS_{max} = 50$ in increments of $CS_{inc} = 5$. This reflects the sizing conventions common in PISA data. School management ranks students by an index R_{sch} comprised of observed discipline and an additional noise, reflecting other factors, such as student aptitude and preferences, taken into consideration when sorting students into classes, so that $R_{sch} \sim N(D_{stu}, \sigma_{rnk} = 0.5e)$. After sorting the students according to R_{sch}, the school then allocates them into classes in order, starting

[12]The inverse-logit transformation was chosen for simplicity, and does not necessarily mimic the exact mapping from students' individual discipline index to their probability of not disrupting the class.

from the largest possible class size ($CS_{max} = 50$).

School management then checks for class discipline $P^{sch}_{class} = \prod_{i=1}^{n_{class}} P^{sch}_i$ (using the discipline levels it observes, where $P^{sch}_i = invlogit(D^{sch}_i)$). Class discipline has to meet or exceed a country-specific threshold:

$$P^{thresh}_{cnt} = \frac{2}{3} * \bar{P}_{cnt}^{\frac{CS_{min}+CS_{max}}{2}} + \frac{1}{3} * 0.5 \tag{4}$$

If class discipline is above the threshold, the class is finalized and the school moves on to the next class. If not, class size is lowered by $CS_{inc} = 5$ and the process is repeated until the minimum class size of $CS_{min} = 10$ is reached. Notice that the threshold is set so that high discipline countries will have both larger classes and stronger classroom discipline, and that on average, the most disciplined students will be placed in the largest classes.

The smallest class size (10) is dropped, to deal with an artifact of the algorithm, in which small classes sometimes enjoy a high level of class discipline, due to the large drop from 15 to 10 students. In a realistic setting a school would likely disband the smaller class and spread students to larger classes under such circumstances. However, in the interest of simplicity classes of size 10 (a total of 1,120 synthetic students) are regarded as outliers and dropped.

For each country, the quality of $\lceil \frac{N_{stu}}{CS_{min}} \rceil = 50$ (number of students per country, divided by the minimal class size, rounded up) prospective teachers is drawn from a Normal distribution $TQ \sim N(\mu_{TQ} = 100, \sigma_{TQ} = 15)$ (identical distribution for all countries). Schools' information on teacher quality is noisy, so that $TQ^{sch} \sim N(TQ, \sigma^{sch}_{TQ} = 0.1\sigma_{TQ})$. Prospective teachers are sorted by observed quality and assigned to classes, where the better teachers are allocated to the larger classes.[13] All countries have a similar number and quality of teachers available. However, a country with large classes will need fewer teachers, and will therefore enjoy higher teaching quality on average.

Once students and teachers are allocated to classes, educational outcomes are generated through the education production function introduced in Equation 1:

$$E_{i,j} = TQ^{\beta_0}_j P^{\beta_1}_i P^{\beta_2}_j \tag{5}$$

where $\beta_0 = 1$, $\beta_1 = 1$ and $\beta_2 = 1$, for simplicity.[14] i indicates the student and j the class he/she belong

[13] σ^{sch}_{TQ} can be viewed as noise to school information on teacher quality, but also as the effect of different considerations in teacher allocation, such as status/seniority, specialization etc.

[14] This simulation favors simplicity and clarity over accurately matching the moments in the data.

to. E is normalized to an average of 500 and a standard deviation of 100, to mimic PISA methodology. The simulated test picks up the students' educations outcomes with a measurement error, so that $\hat{E}_{i,j} \sim N(E_{i,j}, \sigma_E = 25)$.

Individual discipline is also measured with noise: $D_i^{test} \sim N(D_{stu}, \sigma_{test} = 0.2e)$. An additional layer of noise is added at the class level: $D_j^{test} \sim N(logit(P_j), 3 * \sigma_{test} = 0.6e)$. This reflects class/school-level measurement noise, as well as the fact that this paper uses different indicators for student-level and class-level discipline, namely truancy and tardiness for student-level discipline and classroom disciplinary atmosphere for class-level discipline.

D_i^{test} and D_j^{test} are normalized to the 0.01-1 range.

7.2 Synthetic Data Analysis

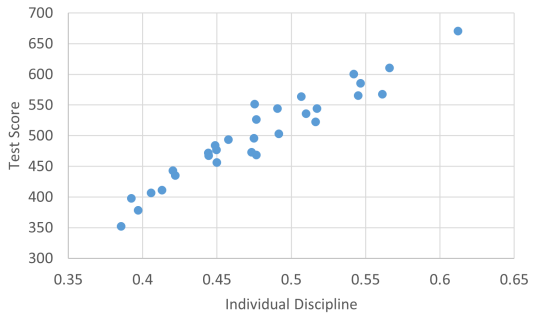
Having Generated synthetic students and teachers and sorted them into classes, correlations in synthetic data can now be compared to those in real PISA data. Table A4 in the appendix presents the breakdown of synthetic data by class size in terms of frequency (total number of students per class size), D_{stu} (normally-distributed individual student discipline), D_{sch} (school-perceived individual student discipline), TQ^{sch} (school-perceived teacher quality) and test scores.

As Figure 9 demonstrates, the synthetically generated data successfully replicates the positive correlations between individual student discipline, classroom discipline, class size and test scores in the PISA data across countries.

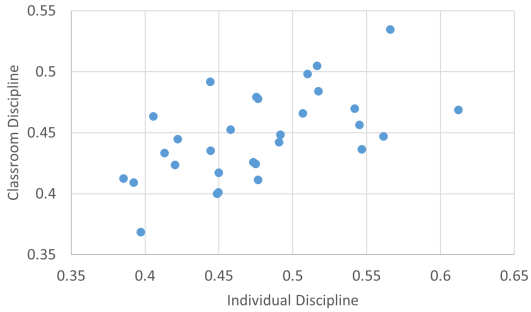
Recall that each country draws a different discipline country average (D_{cnt}) around which its students' discipline level is centered, and which determines the class discipline threshold for the country. Therefore, countries with a higher discipline draw have on average more disciplined students and larger and more disciplined classes. Due to the larger class size they require less teachers, and are able (on average) to select the better ones out of the pool of prospective teachers. The better teachers are assigned to the larger classes (on average), strengthening the positive correlation between class and test scores.

The synthetic data has the added benefit of observing teaching quality. Figure 9 presents the correlations between school measured teaching quality (TQ^{sch}), individual discipline, test scores, class size and class discipline.

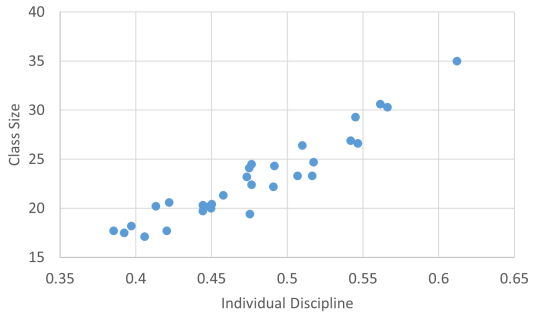
Table 4 presents regressions on synthetic data, correlating synthetic scores with teaching quality, discipline measures and class size dummies.



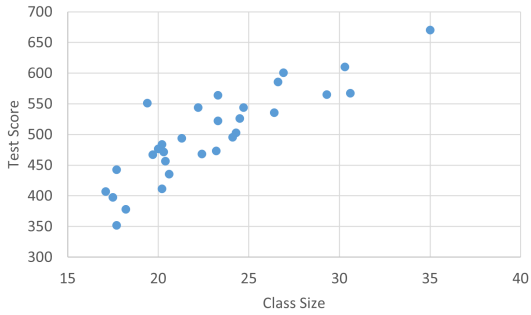
(a) Individual Discipline vs. Test Scores



(b) Individual Discipline vs. Classroom Discipline



(c) Individual Discipline vs. Class Size



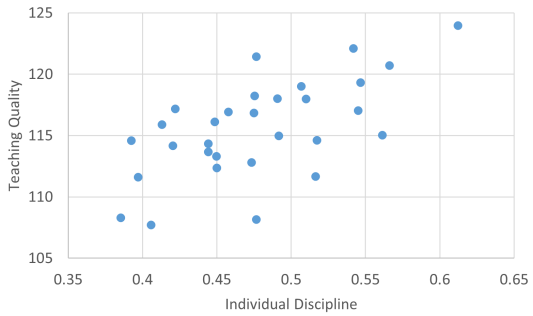
(d) Class Size vs. Test Scores

Figure 8: Correlations in synthetic data, without teaching quality, for 30 synthetic countries

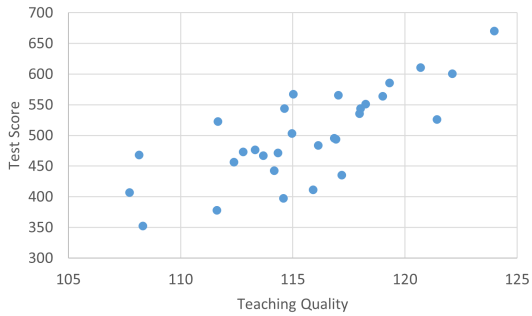
Column (1) presents the full specification, including TQ^{sch} (teaching quality as measured by schools). This specification is close to the data generating process and by construction does not suffer from omitted variable bias. As can be expected, the coefficients are all highly statistically significant and in the correct sign. The class size fixed effects, compared to the baseline of class size 15, are negative and decreasing (increasing in size) in class size. That is, once teaching quality is accounted for, and for given levels of discipline, increasing class size decreases scholastic achievements.

Columns (2)-(6) drop teaching quality, and thus arguably suffer from the same omitted variable bias as the actual PISA data. Notice that these regressions are specified similarly to Table 3.

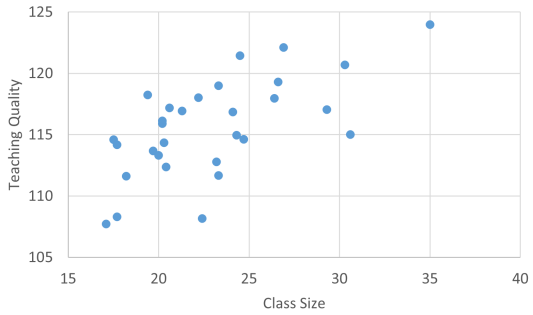
Column (2) keeps only class size fixed effects, similar to Table 3 Column (1). The class size fixed effects are now positive and increasing in class size. This makes sense, since being in a large class is correlated to higher individual discipline and better teachers. Notice that the increase in class size peaks at 40, similar to the real data, where it peaks at class size 36-40. For the real data it can be argued that large classes can sometimes be the result of funding / staffing difficulties, obfuscating the positive effect of bigger classes. This line of argument does not hold for synthetic data. This result can attributed to the noise in the sorting



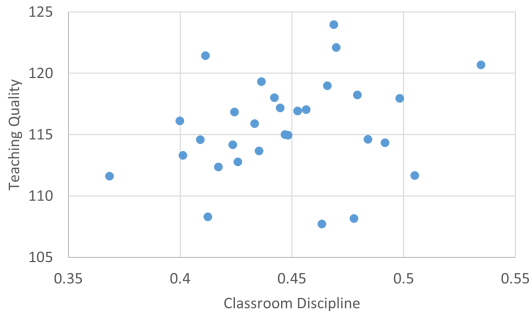
(a) Teaching Quality vs. Individual Discipline



(b) Teaching Quality vs. Test Scores



(c) Teaching Quality vs. Class Size



(d) Teaching Quality vs. Classroom Discipline

Figure 9: Correlations in synthetic data, with teaching quality, for 30 synthetic countries

process: schools measure class discipline using individual student discipline which they observe with noise. Thus, it makes sense that for the larger classes the students' discipline is observed with an upward bias. In other words, conditional on being selected for a large class, the school-perceived individual discipline will be on average higher than the actual individual student discipline. Table A4 shows this indeed the case: the school-perceived individual student discipline (D_{sch}) is larger (on average) than actual individual student discipline (D_{stu}) for all class sizes greater than 15, and the gap is especially large for class sizes 45 and 50.

Column (3) adds P_i (individual discipline). It is positive and highly statistically significant, similar to Table 3 Column (2). Compared to Column (1) the coefficient of individual discipline is larger, and therefore biased upwards (as more disciplined students are likely to be in larger classes, which in turn are likely to have better teachers). Also similar to Table 3 Column (2) is the fact the class size fixed effects drop in size with the addition on individual discipline, since the two variables are correlated.

Column (4) adds $\bar{P}_j$ (class mean individual discipline). It too is positive and highly statistically significant, similar to Table 3 Column (3). Compared to Column (1) the coefficient of class mean individual discipline is also biased upwards. Also similar to Table 3 Column (3) is the drop in size of the coefficient of P_i, as

Table 4: Regressions on synthetic data: student scores vs. teaching quality and discipline measures

	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{TQ_j^{sch}}$ (teaching quality)	8.199 (.106)***					
P_i (ind. disc.)	89.730 (3.987)***		121.972 (4.848)***	111.806 (4.774)***	111.690 (4.769)***	110.215 (4.755)***
$\bar{P}_j$ (class mean ind. disc.)	16.089 (.812)***			22.703 (.969)***	22.584 (.968)***	22.392 (.965)***
σ_{P_j} (class SD ind. disc.)	-33.048 (4.003)***				-26.358 (4.794)***	-24.009 (4.783)***
P_j (class discipline)	40.682 (2.953)***					35.239 (3.529)***
Class Size						
20	-52.233 (1.500)***	5.977 (1.635)***	3.391 (1.602)**	1.961 (1.572)	2.004 (1.571)	3.403 (1.572)**
25	-101.022 (2.008)***	20.198 (1.696)***	15.440 (1.669)***	11.478 (1.646)***	11.265 (1.644)***	11.896 (1.640)***
30	-141.843 (2.831)***	43.906 (2.103)***	37.076 (2.074)***	29.583 (2.060)***	30.127 (2.060)***	31.332 (2.056)***
35	-151.010 (3.282)***	57.520 (2.518)***	49.456 (2.483)***	42.537 (2.453)***	42.772 (2.451)***	45.827 (2.461)***
40	-166.000 (4.079)***	90.525 (3.317)***	79.021 (3.276)***	65.720 (3.263)***	66.239 (3.261)***	67.963 (3.254)***
45	-195.543 (5.151)***	62.427 (4.972)***	52.008 (4.880)***	33.562 (4.851)***	33.790 (4.846)***	43.141 (4.918)***
50	-208.662 (5.989)***	71.435 (6.147)***	57.979 (6.035)***	41.405 (5.961)***	43.238 (5.964)***	47.204 (5.956)***
Const.	-480.441 (11.786)***	482.040 (1.231)***	427.141 (2.492)***	339.550 (4.467)***	359.276 (5.726)***	342.257 (5.955)***
Obs.	13880	13880	13880	13880	13880	13880
R^2	.723	.566	.585	.601	.602	.605

Notes: Synthetic country fixed effects included.
*Significant at 10%; **significant at 5%; ***significant at 1%

the student's level of discipline is correlated to the discipline levels of classmates. Class size fixed effects drop in size, whereas there is no similar drop in Table 3 Column (3). This may indicate that the correlation between class size and class mean individual discipline is weaker in PISA data compared to the synthetic data.

Column (5) adds σ_{P_j} (class SD individual discipline). Similar to Table 3 Column (4), it is negative and statistically significant. However, unlike in the actual PISA data, the addition of σ_{P_j} does not lead to a drop in the size and statistical significance of $\bar{P}_j$. Also of interest is the fact that the σ_{P_j} coefficient in Column (5) is smaller in size (less negative) compared to Column (1).

Column (6) adds P_j (class discipline). Similar to Table 3 Column (5), it is positive and highly statistically significant. However, the inclusion of P_j does not reduce the size and statistical significant of the $\bar{P}_j$ and σ_{P_j} coefficients. In addition, compared to Column (1), the P_j coefficient in Column (6) is smaller. The omission teaching quality does not create an upward bias in this case.

Overall, the similarities between Table 4 and Table 3 illustrate that the theoretical model presented in Section 3 offers a viable explanation for the correlations seen in PISA data.

8 Discussion and Conclusion

This paper establishes a theoretical 'education production' framework, which predicts student discipline, class size and teaching quality and scholastic achievement to be positively correlated. It then proceeds to use classroom disciplinary atmosphere and truancy and tardiness survey data from PISA, the OECD Programme for International Student Assessment, to construct novel empirical measures of student discipline (based on truancy and tardiness data) and classroom discipline (based on classroom disciplinary atmosphere data). In line with the predictions of the model, these discipline measures are shown to be strongly correlated with student achievement, as measured by the PISA math score, where individual discipline, average class student discipline and classroom discipline are positively correlated with scores, whereas the standard deviation of class student discipline is negatively correlated with scores.

Importantly, while student discipline is correlated with class size, (within-country) class discipline is not, suggesting that students are sorted into classes and class sizes are determined in an effort to equalize discipline levels across classes, and/or meet some common country-level discipline threshold, as the model predicts. Large classes, up to a certain (normative) size, are correlated with both higher student discipline level and higher test scores, suggesting that large, normative classes get allocated better teaching resources, also in line with the suggested model. On a systemic level, high discipline allows efficient learning in large classes, enabling schools to be more selective and attract a higher caliber of teachers.

The difficulty of observing teaching quality is a major hurdle in estimating education production. Lacking a strong identification mechanism, this paper uses synthetic data to generate students and teachers and assign them to classes, demonstrating the ability of the suggested model to replicate the main empirical relations, and examining the bias created by omitting teaching quality from the empirical analysis.

Results presented in this paper clearly point to student discipline as a crucial factor in explaining gaps in scholastic performance between students, schools and countries. For example, Israel and Hong-Kong have similarly large classes but are on opposite sides of the spectrum when it comes to student discipline. This paper argues that this is a major reason for the massive gap in their average PISA scores.

There is much that remains to be studied regarding this topic, and in particular regarding the interaction between discipline, class formation and teacher quality. Future research could further understanding of

these links by exploiting exogenous measures affecting discipline, class-size and/or teacher allocation, as well as by refining the theoretical model presented in this paper to better match specific educational systems.[15] With regard to teaching quality, while this paper focused on the demand side (the ability of school systems with larger fewer classes to be more selective in the hiring of teachers) the effect of discipline and class size on the supply side (the attractiveness of the teaching profession) is of great interest.

In conclusion, this paper sheds light on how discipline, class size, and teacher quality interact to produce educational outcomes. It suggests that policies to improve classroom discipline and reduce variance in student discipline could yield substantial dividends by enabling effective learning in larger classes and a better teaching force. Ultimately, understanding these interactions is essential for efficient educational resource allocation.

[15]In addition, were PISA to add a module testing teachers alongside their students, such data could prove useful for future research.

References

Angrist, Joshua and Victor Lavy, "Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement," *The Quarterly Journal of Economics*, 1999, *114* (2), 533–575.

Arum, Richard and Melissa Velez, *Improving learning environments: School discipline and student achievement in comparative perspective*, Stanford University Press, 2012.

Asadullah, M. Niaz, Alain Trannoy, Sandy Tubeuf, and Gaston Yalonetzky, "Measuring educational inequality of opportunity: pupils effort matters," *World Development*, 2021, *138.*

_ **, Liyanage Devangi H. Perera, and Saizi Xiao**, "Vietnams extraordinary performance in the PISA assessment: A cultural explanation of an education paradox," *Journal of Policy Modeling*, 2020, *42* (5), 913–932.

Bosworth, Derek, "Truancy and Pupil Performance," *Education Economics*, 1994, *2* (3), 243–264.

Goodman, Joshua, "Flaking Out: Student Absences and Snow Days as Disruptions of Instructional Time," *NBER Working Paper*, 2014, (20221).

Hanushek, Eric A., "The Failure of Input-Based Schooling Policies," *The Economic Journal*, 2003, *113*, F64–F98.

_ **and Ludger Woessmann**, "The Economics of International Differences in Educational Achievement," in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds., *Handbook of the Economics of Education*, Vol. 3, Amsterdam: North-Holland, 2011, pp. 89–200.

_ **and Steven G. Rivkin**, "Teacher Quality," in Eric A. Hanushek and Finis Welch, eds., *Handbook of the Economics of Education*, Vol. 2, Amsterdam: North-Holland, 2006.

_ **, John F. Kain, and Steven G. Rivkin**, "Why Public Schools Lose Teachers," *Journal of Human Resources*, 2004, *39* (2), 326–354.

Hoxby, Caroline M., "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *The Economic Journal*, 2000, *113*, F64–F98.

Jepsen, Christopher and Steven Rivkin, "Class size reduction and student achievement the potential tradeoff between teacher quality and class size," *Journal of human resources*, 2009, *44* (1), 223–250.

Krueger, Alan B., "Experimental Estimates of Education Production Functions," *The Quarterly Journal of Economics*, 1999, *114* (2), 497–532.

_ , "Economic Considerations and Class Size," *The Economic Journal*, 2003, *113* (485), F34–F63.

Lazear, Edward P., "Educational Production," *The Quarterly Journal of Economics*, 2001, *116* (3), 777–803.

Perera, Liyanage Devangi H. and M. Niaz Asadullah, "Mind the gap: What explains Malaysia's underperformance in Pisa?," *International Journal of Educational Development*, 2019, *65* (C), 254–263.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez, "When Students Don't Care: Reexamining International Differences in Achievement and Student Effort," *Journal of Human Capital*, 2019, *13* (4), 519–552.

A Additional Data

Table A1: Median Class Size, by Country and Year

Country	2012	2015	2018	2022	Total
AUS	21-25	21-25	21-25	21-25	21-25
AUT	21-25	21-25	21-25	21-25	21-25
BEL	16-20	16-20	16-20	16-20	16-20
CAN	26-30	21-25	NA	NA	26-30
CHE	16-20	16-20	16-20	16-20	16-20
CZE	21-25	21-25	21-25	21-25	21-25
DEU	26-30	21-25	21-25	21-25	21-25
DNK	21-25	21-25	21-25	21-25	21-25
ESP	21-25	26-30	26-30	21-25	21-25
EST	21-25	21-25	21-25	21-25	21-25
FIN	16-20	16-20	16-20	16-20	16-20
FRA	26-30	26-30	26-30	26-30	26-30
GBR	26-30	21-25	21-25	26-30	21-25
GRC	21-25	21-25	21-25	21-25	21-25
HKG	31-35	31-35	26-30	26-30	26-30
IRL	21-25	21-25	21-25	21-25	21-25
ISR	31-35	26-30	26-30	31-35	26-30
ITA	21-25	21-25	21-25	21-25	21-25
JPN	36-40	36-40	36-40	36-40	36-40
KOR	31-35	31-35	26-30	21-25	26-30
NLD	26-30	26-30	26-30	21-25	26-30
NOR	26-30	21-25	21-25	NA	21-25
NZL	26-30	21-25	21-25	21-25	21-25
PRT	21-25	21-25	26-30	21-25	21-25
SGP	36-40	36-40	31-35	31-35	36-40
SVK	21-25	21-25	16-20	16-20	21-25
SVN	21-25	21-25	21-25	21-25	21-25
SWE	21-25	21-25	NA	21-25	21-25
TAP	36-40	36-40	36-40	31-35	36-40
USA	26-30	26-30	26-30	26-30	26-30
Total	21-25	21-25	21-25	21-25	21-25

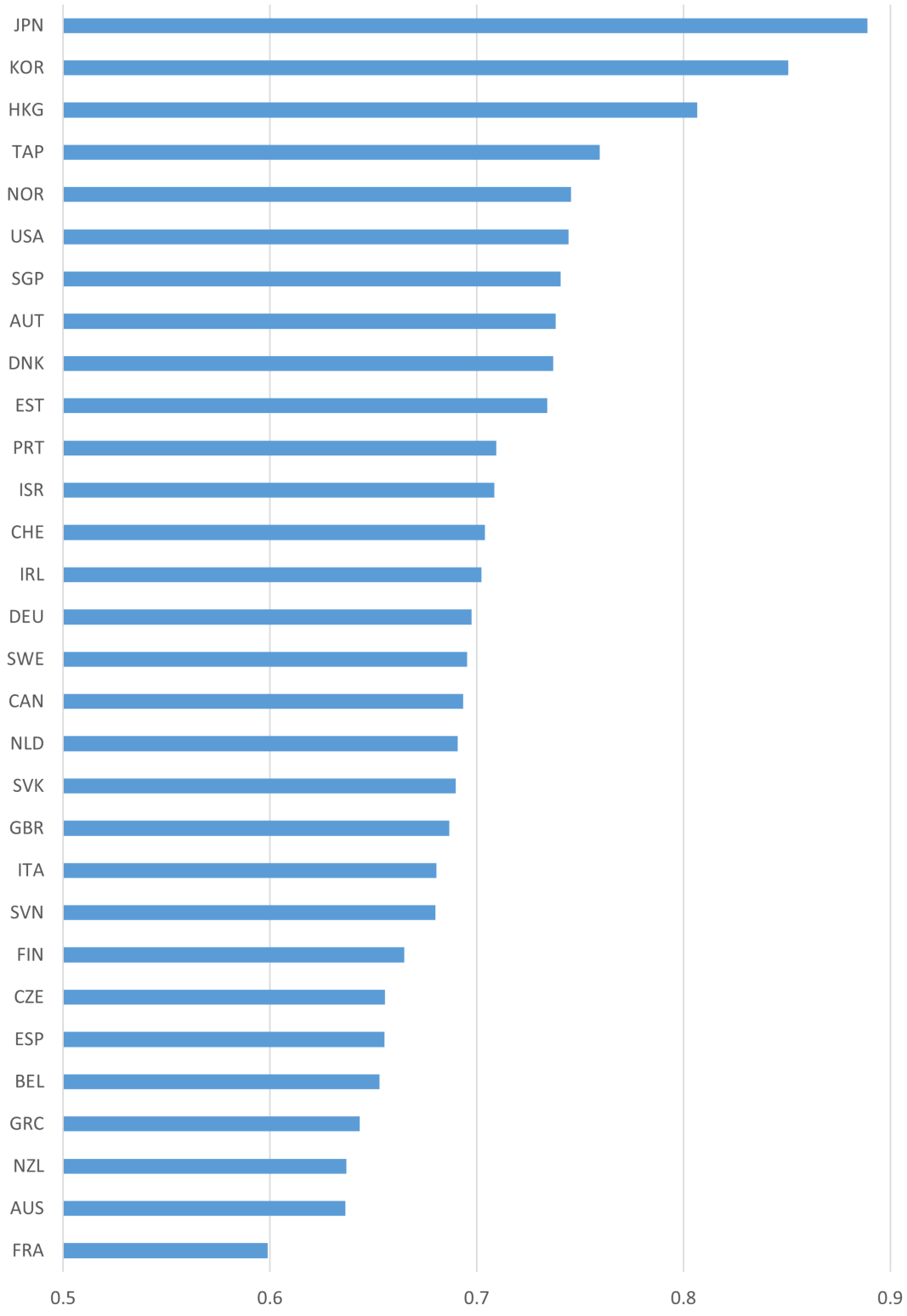


Figure A1: Classroom discipline index (2012, 2015, 2018 and 2022 round averages)

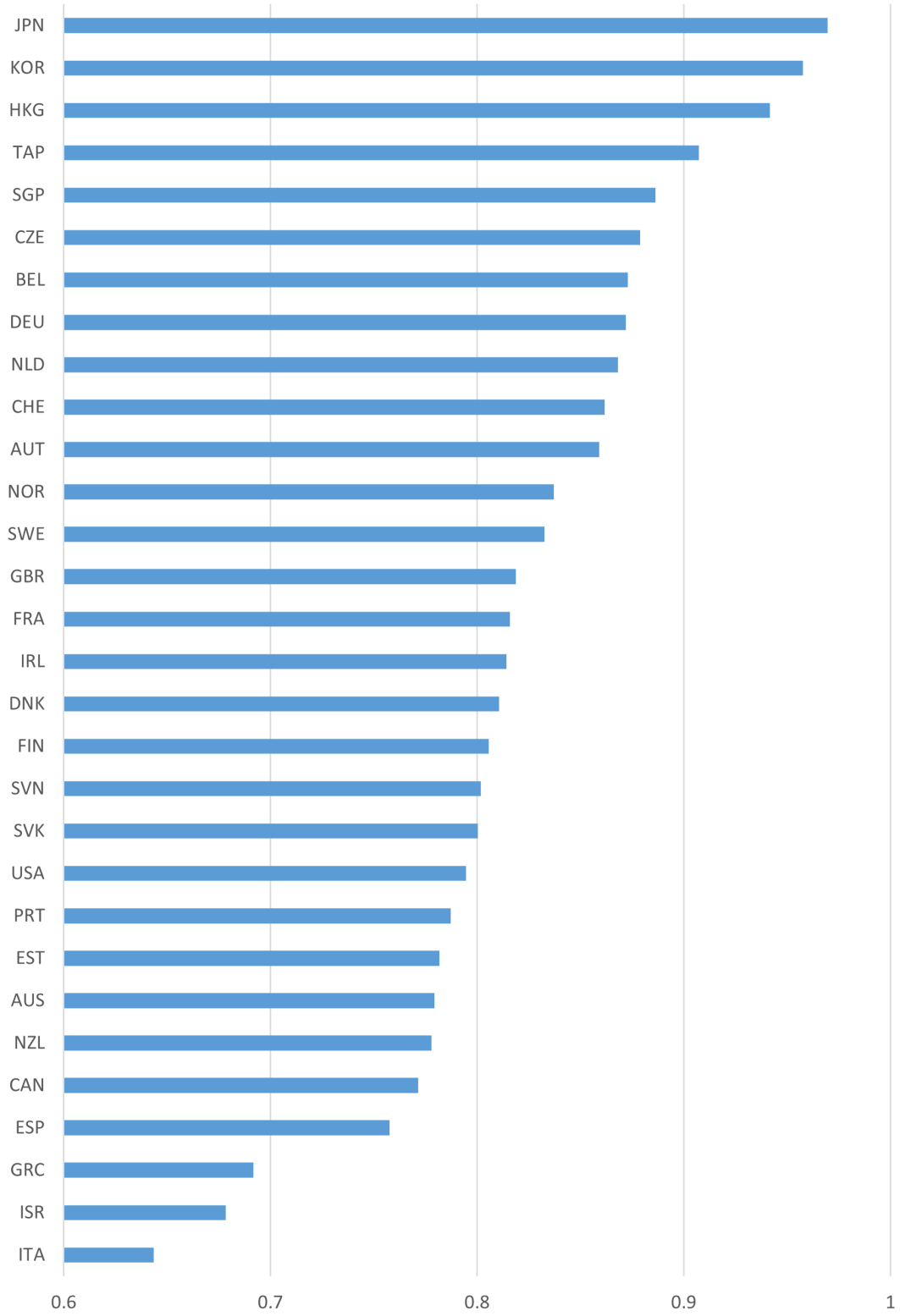


Figure A2: Individual discipline index (2012, 2015, 2018 and 2022 round averages)

Table A2: Regressions on student-level PISA math scores

	All Class Sizes	**15 or less**	**16-20**	**21-25**	**26-30**	**31-35**	**36-40**	**41-45**	**46-50**	**51+**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
P_i (ind. disc.)	54.429	55.258	51.732	52.052	56.317	55.283	73.260	75.000	57.294	47.487
	(3.802)***	(3.676)***	(4.368)***	(4.606)***	(3.585)***	(8.186)***	(8.795)***	(4.377)***	(8.412)***	(9.225)***
$\bar{P}_j$ (class mean ind. disc.)	49.170	94.122	88.738	68.573	30.164	-49.996	47.512	-4.934	35.187	122.246
	(36.579)	(25.118)***	(34.495)**	(43.065)	(29.587)	(48.206)	(55.521)	(60.589)	(69.004)	(75.060)
σ_{P_j} (class SD ind. disc.)	-77.079	-38.351	-29.253	-41.869	-81.113	-192.444	-158.346	-180.290	-219.809	56.314
	(35.061)**	(21.686)*	(33.733)	(37.343)	(32.432)**	(39.770)***	(69.685)**	(80.751)**	(39.751)***	(74.974)
P_j (class discipline)	71.494	63.982	69.950	66.281	66.432	69.488	94.164	88.446	56.943	115.726
	(5.616)***	(8.478)***	(7.737)***	(6.153)***	(6.401)***	(17.442)***	(15.235)***	(45.339)*	(18.208)***	(39.769)***
Class size										
16-20	8.140									
	(1.982)***									
21-25	14.617									
	(2.926)***									
26-30	21.999									
	3.365)***									
31-35	31.387									
	(4.810)***									
36-40	37.284									
	(5.084)***									
41-45	34.082									
	(7.574)***									
46-50	16.753									
	(5.211)***									
51+	8.928									
	(2.497)***									
Const.	246.722	275.341	181.086	256.530	335.299	310.703	121.794	-261.517	167.303	-124.846
	(56.617)***	(60.627)***	(59.988)***	(69.036)***	(69.345)***	(68.227)***	(91.344)	(184.931)	(127.776)	(96.243)
Obs.	740340	23179	115382	260376	212263	61125	43140	9572	4877	10426
R^2	.274	.242	.235	.245	.248	.328	.291	.423	.344	.288

Notes: Controls for year, country, country GDP per capita, country share of population under age 15, maternal and paternal education and share of schoolmates whose parents have post-secondary education. Errors clustered by country.
*Significant at 10%; **significant at 5%; ***significant at 1%

Table A3: Student-level PISA math scores with interactions

	X P_i	X $\bar{P}_j$	X σ_{P_j}	X P_j
	(1)	(2)	(3)	(4)
P_i (ind. disc.)	55.831 (4.718)***	54.427 (3.802)***	54.427 (3.800)***	54.423 (3.800)***
$\bar{P}_j$ (class mean ind. disc.)	50.834 (36.924)	52.611 (35.126)	53.924 (36.041)	50.854 (36.542)
σ_{P_j} (class SD ind. disc.)	-75.165 (35.204)**	-71.968 (34.939)**	-50.062 (32.116)	-74.175 (34.564)**
P_j (class discipline)	71.538 (5.630)***	71.698 (5.665)***	71.464 (5.610)***	68.271 (8.948)***
Class Size				
16-20	12.790 (2.575)***	18.234 (6.240)***	5.510 (4.457)	6.697 (4.914)
21-25	18.509 (3.611)***	21.566 (9.524)**	14.429 (5.034)***	15.879 (7.289)**
26-30	21.128 (4.137)***	15.544 (13.506)	31.067 (5.716)***	20.476 (7.897)***
31-35	33.127 (11.035)***	31.331 (30.049)	42.012 (6.831)***	37.653 (22.777)*
36-40	14.557 (16.023)	-18.839 (58.551)	59.961 (9.012)***	-8.523 (21.053)
41-45	-11.807 (9.212)	-124.968 (58.710)**	75.545 (7.879)***	-42.294 (22.789)*
46-50	5.611 (7.430)	-35.408 (16.949)**	47.563 (6.678)***	-11.761 (14.782)
51+	10.835 (3.847)***	-11.146 (27.941)	18.076 (15.600)	-25.853 (25.239)
Class Size Interactions				
16-20 interaction	-5.765 (2.827)**	-12.541 (8.072)	11.777 (15.946)	2.063 (6.520)
21-25 interaction	-4.799 (3.437)	-8.542 (12.74)	1.477 ((13.531)	-1.820 (9.746)
26-30 interaction	1.138 (4.548)	8.239 (17.517)	-39.818 (19.480)***	2.280 (10.650)
31-35 interaction	-1.904 (12.114)	.881 (35.163)	-47.716 (31.009)	-8.092 (26.758)
36-40 interaction	25.746 (16.835)	64.521 (65.374)	-126.047 (48.660)***	59.275 (26.175)**
41-45 interaction	51.152 (11.123)***	177.649 (65.528)***	-259.051 (51.669)***	98.288 (31.548)***
46-50 interaction	13.580 (11.476)	63.390 (21.189)***	-144.979 (26.455)***	40.051 (23.691)*
51+ interaction	-2.228 (5.393)	26.559 (37.581)	-38.255 (61.710)	50.685 (37.301)
Const.	245.571 (56.346)***	247.797 (56.724)***	247.204 (57.084)***	245.072 (55.626)***
Obs.	740340	740340	740340	740340
R^2	.274	.275	.275	.275

Notes: Controls for year, country, country GDP per capita, country share of population under age 15, maternal and paternal education and share of schoolmates whose parents have post-secondary education. Errors clustered by country.
*Significant at 10%; **significant at 5%; ***significant at 1%

Table A4: Synthetic data summary Statistics by Class Size

Class Size	Freq.	D_{stu}	D_{sch}	TQ^{sch}	Test Score
15	3,105	3.96	3.89	106.38	456.92
20	3,720	4.15	4.17	112.74	475.85
25	3,425	4.32	4.38	119.00	500.98
30	1,650	4.49	4.58	125.79	533.42
35	1,085	4.77	4.84	127.61	588.28
40	520	4.92	4.98	132.12	621.20
45	225	5.13	5.29	130.42	635.89
50	150	5.34	5.49	132.41	663.76
Total	13,880	4.30	4.32	116.80	503.52

PUBLICATIONS

Class Discipline, Class Size and Scholastic Achievement Across Countries: A Theoretical and Empirical View of Educational Production

Working Paper No. WP/2026/105