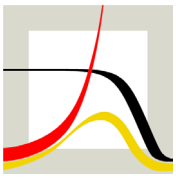

Max-Planck-Institut für demografische Forschung

Max Planck Institute for Demographic Research

Konrad-Zuse-Strasse 1 ▪ D-18057 Rostock ▪ Germany ▪ Tel +49 (0) 3 81 20 81 - 0 ▪ Fax +49 (0) 3 81 20 81 - 202 ▪ www.demogr.mpg.de

MPIDR Working Paper WP 2020-005 I February 2020
https://doi.org/10.4054/MPIDR-WP-2020-005

A New Perspective On The International Achievement Gap: Is Academic Autonomy Good For Everyone?

Jorge Cimentada I cimentada@demogr.mpg.de

This working paper has been approved for release by: Emilio Zagheni (sekzagheni@demogr.mpg.de), Head of the Laboratory of Digital and Computational Demography.

A New Perspective On The International Achievement Gap: Is Academic Autonomy Good For Everyone?

Jorge Cimentada

01 November, 2019

Abstract

There is a growing literature and interest on the study of the cognitive achievement gap between the top and bottom SES groups. Amidst public concern for this distancing between social classes, researchers have been unable to find an adequate explanation for the increasing cross-country inequality. In this paper, I argue that we need to refocus our efforts towards understanding better what correlates with the academic performance of both SES groups separately. By shifting attention to the amount of school autonomy that different schools have, I show that school autonomy over academic content, courses and text books is associated with a decrease of test scores of nearly .4 standard deviations for the bottom 10% performers in mathematics and literacy – a whole grade's worth of knowledge. I show that this relationship holds under several specifications, including fixed effect models. In contrast, the same relationship turns positive when relating to the top 10% of students but it's much weaker than for the bottom performers. These results point out that perhaps an explanation to the changing gaps is not symmetrical between groups but rather group specific. The importance of understanding what affects separate SES groups is paramount to understanding the achievement gap and these preliminary results can have important implications in policy making as they speak directly to education policy makers trying to fine tune the autonomy measures of their country.

1 Background

1.1 The National And International SES Achievement Gap

The cognitive achievement gap between the most and least advantaged children is growing steeply over time (Chmielewski 2019). This pattern of increasing inequality between SES groups has already been investigated in countries such as the United States (Reardon 2011) and other developed countries (Bradbury et al. 2015) but there is still much that is not known on the causes and correlates of the evolution of the SES achievement gap. This relationship is of particular importance for policy making as it can help understand the mechanisms under which different SES groups are faring better or worse under increasing economic and social inequality.

Most of the literature on achievement gaps has concentrated on comparing the magnitude of the differences between and within countries on the cognitive achievement gap (Micklewright and Schnepf 2006; Reardon 2011; Vandenberge 2006). This literature has shown that, similarly to the income distribution gap (Alvaredo et al. 2017; Milanovic 2016), the SES cognitive achievement gap is also drastically different between countries. For example, evidence points out that the United States has the biggest achievement gap of all countries with a total of 1.25 standard deviations and Iceland has the narrowest gap with a gap centered at around 0.75 standard deviations (Reardon and Portilla 2016). Amidst this growing concern, researchers have tried to focus on trying to explain why this gap is so different between countries. In particular, many have payed attention to inequality indicators as there seems to be some relationship with income inequality (Chmielewski and Reardon 2016) as well as with social inequality (Duru-Bellat and Suchaut 2005). Others have discarded explanations such as the curricular setup of a country and the level of segregation given the weak correlations to the cross-country differences (Duru-Bellat and Suchaut 2005).

However, nearly all of the previous studies focus on the cross-national comparison of the achievement gap. In recent years attention has shifted to studying the trends in achievement gap rather than solely focusing on the static SES gaps. The first attempts to study the *evolution* of the achievement gap was done in the United States and it documented that the gap in cognitive abilities between high-SES and low-SES children has been widening over the years (Reardon 2011). Using over 60 years of data on educational testing surveys,

Reardon (2011) found that not only has the cognitive gap between the 90th income percentile and the 10th income percentile grown over time, but it has grown faster and to be wider than the highly contested white-black achievement gap (Magnuson and Waldfogel 2008). According to his findings, these gaps have actually reversed and we find that the income achievement gap is nearly twice as large as the black-white achievement gap (quite the opposite to 20 years back). Interestingly, the widening of the achievement gap has been paralleled by a growth of income inequality, which may be telling. Reardon (2011) offers several possible links, with the most reasonable being that family investment patterns have changed so that high income families now invest more resources on their children. The explanation lies in the fact that increasing income became more strongly correlated with other positive family traits related to time allocation and welfare services.

In a follow-up study, Reardon and Portilla (2016) uncovered a reversal of the trend. The follow-up study concentrated solely on kindergarten children in the U.S. for the years 1998, 2006 and 2010. They found that the 90th/10th income gap in readiness closed modestly. Furthermore, using data from fall and spring in the same kindergarten year, they calculated that the gap narrowed at a rate of 0.01 and 0.008 SD per year for mathematics and literacy between 1998 and 2010. They also calculated the same changes for a number of personality traits such as self-control and externalizing behavior and found similar results. In contrast, Reardon (2011) finds that in a 30-year span the gap was systematically increasing at a rate of 0.02, something reasonably close to the previous estimates. Their results not only hold for the income achievement gap, but they also found a decline in the white-hispanic gap (although not for the white-black gap). The reasons why the authors find a reversal in the trend could be numerous and should be studied closely. They discuss a number of country-level indicators to explain this change and suggest that the reversal is likely due to the high increase of preschool enrollment from the low SES group. They build on their previous argument by suggesting that in this same period (1998 - 2010) the income achievement gap in early schooling enrollment decreased substantially. Their conclusions, although suggestive, are speculative and have no *empirical support* which is why this is still an open question.

There have been other attempts to explain the achievement gaps with indicators such as economic inequality (Dupriez and Dumay 2006), the difference in schooling hours and the tracking system (Duru-Bellat and Suchaut 2005; Dupriez and Dumay 2006), home and family factors (Marks, Cresswell, and Ainley 2006) and expanding school access (Chmielewski 2019). The work of Dupriez and Dumay (2006) explored the relationship between achievement gaps and economic inequality but without factoring in the multilevel structure of the students nested into schools. Moreover, it merely correlated achievement gaps with economic inequality. The work of Duru-Bellat and Suchaut (2005)]is more comprehensive as it explores several indicators of the school system, among which is the differentiation structure of the secondary school system (tracking). However, as noted by Reardon, Robinson, and Weathers (2008), *'our understanding of the causes and patterns of these achievement gaps is far from complete'*. For this reason, the review by Van de Werfhorst and Mijs (2010) gains particular relevance because it documents many instances in which tracking explains inequality between schools (one notable example is the work of Dupriez and Dumay (2006) which finds a strong correlation between tracking and achievement gaps).

Motivated by these recent results, other authors have taken this analysis to an international context in order to discover between-country trends. The work of Bradbury et al. (2015) employs a unique comparative analysis of the achievement gap between Australia, United Kingdom, United States and Canada. Their research design is distinctive in that they use longitudinal data from children as early as age 2 and study the evolution of the achievement gap up until age 14 [1]. The core finding of their study is that the American achievement gap is much wider than the gaps in Australia and Canada. They find that once the achievement gap is present in early school entry, it does not seem to narrow or widen much over the life course. In fact, they estimate that the quality of early childhood education can only explain about 30-40% of the high school SES gap. This suggests that once the achievement gap is present before entering school, it carries a social-scar effect.[2] One exception is the UK, which they found to be a country that helps close the gap in

[1] To the best of my knowledge this is not only the first study that uses panel data to study achievement gaps, but to also do it between countries

[2] However, schooling could be preventing the gap from widening even more, and rigorous Randomized Controlled Trials (RCT) show that high quality schooling can indeed help ease the gap, in some instances even close it (Campbell et al. 2002)

early primary years. This can likely be due to the comprehensive schooling and also the public support by the welfare state in dimensions like health and income support.

In a similar line but using data on more than 80 countries, Chmielewski (2019) compares data from 30 international large-scale assessments over 50 years of data to study the global evolution of the achievement gap. Despite big cross-national variation in the evolution of the achievement gap, there is a widespread trend of increasing achievement gaps for the three SES proxies used in the study: parent's education, parent's occupation and books in the household. In fact, the hard numbers point out that the achievement gap between the three indicators increased at a rate of 0.007-0.008 standard deviations per year. These increases add up to a total of .4 standard deviations in the time span of the study, a sizable increase in the gap. It's also reassuring as the average yearly increase of Chmielewski (2019) matches the magnitude of the point estimates from Reardon (2011) and Reardon and Portilla (2016) which are around 0.01 and 0.02.

However, one drawback of these over time cross-country analysis is that they adjust for age as most of the studies they use come from children at different stages in their school trajectory. However, there is evidence which suggests that achievement gaps are indeed exarcebated differently at different time points in the school trajectory (Hanushek and others 2006; Van de Werfhorst and Mijs 2010; Bradbury et al. 2015). This means that most of these studies mask age-specific gaps in exchange for overall yearly gaps for each country separately. Among the few studies which concentrate on age-specific gaps, there are already quite different results from the main ones discussed in Chmielewski (2019) and Broer, Bai, and Fonseca (2019). For example, Reardon and Portilla (2016) and Hanushek et al. (2019) find a decrease in the gap and a static gap when focusing in a narrow set of ages[3].

Age-specific gaps are indeed quite different as they reflect the overall distance between SES groups at the same point in time in different years (achievement gap for 15 year olds in many different years). These age-specific gaps has at least one benefit relative to age pooled trend analysis. It allows to compare students at the same time point, holding constant any differences which come about with age (in pooled analysis, age is controlled for but not all of the associated differences to inequality across the lifespan, such as for example, the stronger effect of inequality at earlier ages (Kulic et al. 2019)). By focusing on one age gap, students are at a point where the intensity of inequality is absored similarly by everyone due to the same age in the life-cycle skill formation (Cunha et al. 2006).

As can be seen from the evidence here, there is no clear consensus on what might be driving these within country changes. But even more, we're far from establishing some general explanations for the between country differences in the gap. Instead of focusing on a single or joint explanation for the evolution of the achievement gap, this paper is interested in exploring the relationship between school level indicators and their relationship to the achievement gap separately by SES achievement groups. More concretely, what is the role of school autonomy in influencing the achievement gap?

1.2 School Autonomy And The Widening Of The Achievement Gap

The concept of school autonomy and it's relationship to increasing equality has been a predominant topic in policy research. Most rigorous studies performed on autonomy interventions document a positive increase in test scores, a decrease in school dropouts and a decrease in grade repetitions among other things (Bruns, Filmer, and Patrinos 2011). However, most of this evidence is concentrated on small-scale pilots and interventions in particular areas of interest (low SES areas in developing countries, for example) (Di Gropello 2006).

The work of Hanushek, Link, and Woessmann (2013) is among the first to document internationally that autonomy seems to be negative for low income countries and positive for high income countries. Focusing on three different types of school autonomy (curricular, personnel, budget), they find consistent evidence that curricular and personnel autonomy seems to be associated with increasing test scores in high income countries and decreasing test scores for low income countries. They argue that the mechanism under which

[3]Hanushek et al. (2019) use the Long-Term Trend National Assessment of Educational Progress, the Main National Assessment of Educational Progress and the Programme for International Student Assessment but focus almost exclusively on children between ages 13 and 15

autonomy can have negative effects is when decision makers are opportunistic in their behaviors but also when decision makers are not well prepared to make these decisions. Ammermüller (2005) actually found that the more autonomy the school has, the more relevance the parent's education and cultural resources gain importance. This means that autonomy is associated with a somewhat negative influence on student's performance as they have to rely more on the input from resources outside of school than on their teacher's and school's resources.

However, the mechanism under which Hanushek, Link, and Woessmann (2013) suggests that autonomy can have negative impact can also be present in developed countries. For example, if the worst performing schools have on average lower quality of teaching, then autonomy can theoretically have negative affects as lower performing teachers could deviate from the validated national curriculum and affect the learning experience. Similarly, teachers could be lowering the academic standards of lower performing students reinforcing their already poor performance through lower goals. In particular, this last mechanism has been discussed in detail in Gamoran and Berends (1987) and Hattie (2002). Conversely, having greater academic autonomy for the good performing students can increase their performance by altering the teaching methods to increase the learning rate of these students.

Most school autonomy related research focuses on between country relationships and the average country performance (Hanushek, Link, and Woessmann 2013; Ammermüller 2005; LeTendre, Hofer, and Shimizu 2003; Stevenson and Baker 1991) without paying attention whether autonomy can have varying effects within a country at the extremes of the achievement gap. School autonomy as a means of school differentiation can have either positive or negative results based on the capacity of the school/teachers to make these decision and on the composition of the students at each school. That is why Van de Werfhorst and Mijs (2010) discuss evidence that standardization in autonomy is often associated with increasing equality in Europe. However, there is no clear evidence on whether autonomy is good for all *within* countries. Moreover, the argument from Hanushek, Link, and Woessmann (2013) is that autonomy is beneficial for high income countries without disaggregating whether it's good for some and bad for others.

This study will focus on studying the relationship between academic autonomy, personnel autonomy and budget autonomy and it's relationship to the highly contested SES achievement gap. In particular, this paper will investigate whether different types of autonomy can have negative associations with the bottom 10% of students (the bottom group of the SES gap) and whether this same relationship is reversed for the top 10% of students (the top group of the SES gap). However, I focus almost exclusively on high income countries to test whether autonomy can also have negative effects within developed economies.

2 Empirical Strategy

2.1 Data

To investigate the above mentioned questions, I will use the Programme for International Student Assessment (PISA). PISA is a survey carried out every three years that aims to evaluate education systems by testing the skills and knowledge of 15-year-old students. Currently, PISA has six waves starting in 2000 up until 2015, where recently, over half a million students were tested in mathematics, literacy and science in over 70 developed/developing countries.

PISA collects data through a two-stage stratified sampling design. With the help of governments, PISA randomly chooses 150 schools in each country, where they then randomly pick thirty 15 year olds to undertake the two hour tests. Together with the subject tests, PISA collects personal information from students, their families and their school environment, that serves as relevant background information that can be matched to the students performance. With the recent inclusion of PISA 2015, these six waves make up a time-series analysis of 15 years, enough to visualize changes in the structure of an educational system. None of the studies cited so far has used the last PISA wave, which was released in December 2016. This chapter takes advantage of these six waves to build a country pseudo-panel, making it possible to study changes in nearly 15 years for 29 countries. In order to maximize country variation, I have included countries which have at least participated in 50% of all waves.

To identify a student's socio economic status I use the composite SES index created by the PISA team. The index of economic, social and cultural status (ESCS) was created on the basis of the following variables: the International Socio-Economic Index of Occupational Status (ISEI), the highest level of education of the student's parents, the PISA index of family wealth (which measures the material wealth of the family), the PISA index of home educational resources; and the PISA index of possessions related to "classical" culture in the family home (mainly about books in the household) (OECD 2002). The variable, aside from capturing all relevant dimensions of SES, such as education, occupation, and material resources, takes care of transforming all mentioned variables into comparable metrics across waves. The ESCS index was derived from a principal component analysis of standardized variables, taking the factor scores for the first principal component as measures of the PISA index of economic, social and cultural status. All countries and economies (both OECD and partner countries/economies) were assigned the same weight in the principal component analysis, while in previous cycles, the principal component analysis was based only on OECD countries. However, for the purpose of reporting, the ESCS scale has been transformed with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries (OECD 2016). To the best of my knowledge this is the first piece of research that uses the newly-released ESCS index (OECD 2016), which was rescaled so that all ESCS indexes are suitable for over-time analysis [4]. In other words, the ESCS index does not need any transformation or coding updates as it is ready for comparison over time.

Aside from SES, the other relevant variables are test scores for mathematics and literacy [5]. PISA does not provide a single test result for each respondent. Instead, it provides a *series* of 'plausible values' that the child could actually score. As explained in the PISA manual (OECD 2012), these are imputed values that resemble individual test scores and have approximately the same distribution as the latent trait being measured (the true distribution of the possible scores a student can achieve) [6].

A more intuitive explanation is this: suppose we have μ_i, the average student test score in mathematics for student i. Instead of estimating μ_i alone, plausible values estimate a distribution of possible μ's for student i, together with the likelihood of each μ_i based on the respondents answers on the test. This is defined as the posterior distributions of μ's for student i. The reason why PISA uses this procedure is because estimating a single number μ_i is plagued with measurement error, among other types of bias (see Wu 2005). The number of plausible values for PISA waves are usually five (although ten for PISA 2015) random draws from this distribution. In practice, each student has 5 scores for each test, which resembles their distribution. Those values are continuous, ranging from 0 to 500, with a mean of 250. However, PISA test scores were scaled to have a mean of 500 and a standard deviation of 100 over students in all OECD countries in the first year of focal testing (e.g. 2000 for mathematics and reading).

As per the independent variables, I will include both student level variable and school level variables. For the student level variables, I will include the gender of each student, their parent's level of education (to control for residual variation within each group of the SES index), the occupation index of their parents (centered at 0), whether they're native students or immigrants, the number of books in their household and the specific school programme they belong to (general, pre-vocational or vocational tracks). At the school level, I include the location of the school (small town, town, large town, city or large city), whether it's a public or private school, the size of the school in number of students (center with a mean of 0) and the percentage of funding that comes from government funding (centered with a mean of 0). In the next section I describe the definition of autonomy measures.

[4] These rescaled indexes can be found at http://www.oecd.org/pisa/data/2015database/ under *Rescaled indices for Trend Analyses*.

[5] The analysis in this paper is mainly concentrated on Mathematics to be able to compare some of the findings with the existent literature which has predominantly focused on this subject. Literacy is used as a second test to check if the results hold. PISA also tests students in Science but since very little research has been done on this subject related to achievement gaps, it was not included in the analysis

[6] It should be noted that PISA has rotating modules for the main subject of that year. This means that the quality of data might be different for the same subject over time

2.2 Methodology And Variables

2.2.1 Definition Of Autonomy

Each school principal in PISA was asked the question of who has the main responsibility for the certain areas in the school. These areas are (1) autonomy over which courses are offered, (2) over the content of the courses, (3) over choosing textbooks, (4) over hiring teachers, (5) over setting their salaries and (6) over budget allocationfor the school. For each of these questions, the principal can answer whether this is the responsibility of the teacher's, the principal's, the school's governing board, a regional or local education authority or a national education authority. Similarly to Hanushek, Link, and Woessmann (2013), I define that a school has autonomy over an area if neither a regional or local authority nor the national education authority has any responsibility in setting these areas. That is, the decision is taken either by a teacher, a principal, or the school's governing board. This definition mimics the already established definition by Hanushek, Link, and Woessmann (2013) and tackles specifically whether the decision is solely responsibility of the school. Table 1 shows the correlation between these autonomy meaures.

Given that course autonomy, content autonomy and text book autonomy are highly correlated, I also include an 'academic content autonomy' index which is just the average between these three. Moreover, I do the same for autonomy over hiring teacher's and setting teacher's salary into the index 'personnel autonomy'. Since school budget autonomy does not correlate with any other variables, I leave it as is.

Table 1: Correlation coefficients between school autonomy measures

	Course	Content	Textbook	Hiring	Salary	School Budget	Academic content	Personnel
Course		.81	.66	.45	.47	.38	.90	.52
Content	.81		.69	.45	.46	.26	.92	.52
Textbook	.66	.69		.45	.24	.39	.88	.42
Hiring	.45	.45	.45		.54	.19	.50	.93
Salary	.47	.46	.24	.54		.16	.43	.80
School Budget	.38	.26	.39	.19	.16		.39	.20
Academic content	.90	.92	.88	.50	.43	.39		.53
Personnel	.52	.52	.42	.93	.80	.20	.53	

Table 1 shows that the index of academic content is highly correlated to autonomy over which courses are offered, which content is offered and autonomy over choosing textbooks (.90, .92 and .88, respectively), which suggests that it represents the concept of academic content autonomy well. Similarly, personnel autonomy has correlations of .93 and .80 with autonomy over teacher hiring and teacher salary, which also validates the index of personnel autonomy. From now on, the main unit of analysis for the autonomy variables will be the index on academic content autonomy, personnel autonomy and budget autonomy. Figure 1 plots the evolution of academic content autonomy and personnel autonomy for all countries over the 6 pisa waves.

There are indeed starking differences between countries. For example, the United States is experiencing a decrease in autonomy over the past 15 years whereas Denmark is witnessing an increase in autonomy in recent years. Hanushek, Link, and Woessmann (2013) discusses in detail some decentralization reforms happening in some of these countries and how they match the currents trends we see in this plot. For example, the increase in autonomy for Germany and the decline in autonomy from the United States reflects underlying reforms to decentralize and centralize decision making respectively. For Germany, the increasing autonomy matches the reforms implemented in North Rhine-Westphalia which give more autonomy to schools over hiring teachers (Hanushek, Link, and Woessmann 2013). Whereas in the United the decreasing trend also reflects the expansion of national level standards from the 'No Child Left Behind' reform. All in all, the evidence points out to some important differences in autonomy between countries accompanied with over time dynamics in decision making.

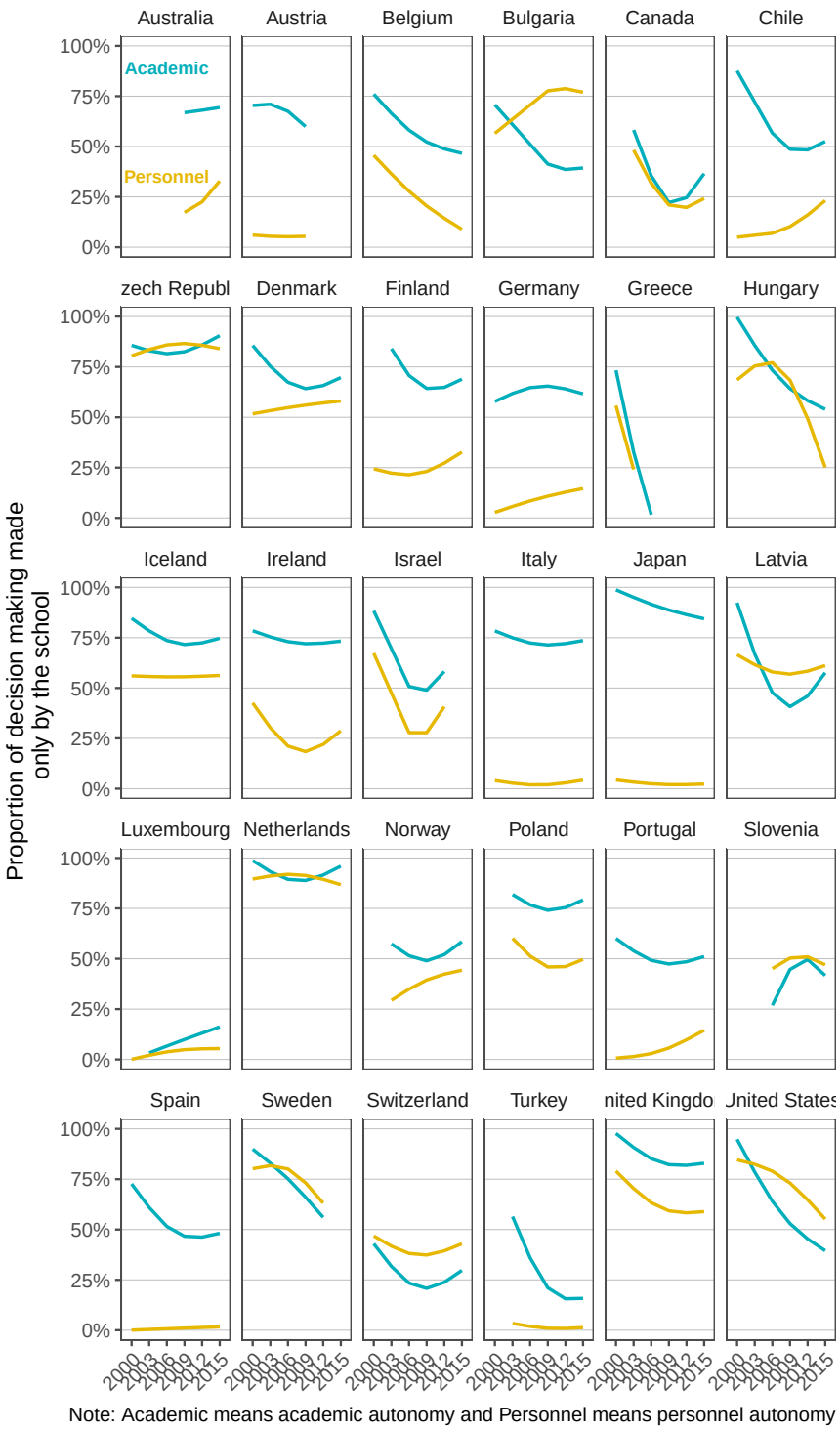

Figure 1: Evolution of school autonomy by country

2.2.2 Model

The model that I will use in the analysis is a cross-classified multilevel model where I allow the country-wave intercepts to vary randomly to account for the clustering of students into each country-year combination. Note that the multilevel approach is just a strategy to account for for the clustering as it adjusts the standard errors appropriately. Yet the specification of a multilevel model is not of particular interest for it's random component as the objective is not to explain cross-country or cross-wave differences in autonomy but rather test whether the relationship between autonomy hold across a sample of countries which are repeated over time, accounting for their nested structure.

The main model can be formally defined as:

$$y_{icy} = \alpha_{cy} + b_1 * autonomy_{cj} + b_n * x_{icj} + \epsilon_{icy} \tag{1}$$

Where y_{icy} is the mathematics test score for student i in country c and year y. α_{cy} is the random intercept for each country/year, b_1 is the beta coefficient for each autonomy measure (models are run separately) and x_{icj} is the vector of covariates used as controls in all models. All models presented are run using the student sample weights to have representative estimates of the populations.

The dependent variable of the analysis will be the standardized mathematics test scores for each student[7]. As mentioned before, PISA does not provide a single achievement indicator. Instead, I calculate the median of all plausible values for each student [8], resulting in one single score.

To standardize the test scores I fit a linear model for each wave, where test scores is regressed on age measured in months (following the same strategy as Reardon (2011) [9]) weighted by the student sample weights from PISA. With the residuals of this adjusted test score, I standardized the metric to solve the problem of comparability over time[10] for all PISA waves. However, another concern is whether test scores measured at different waves have different amounts of measurement error. If that is the case, then the amount of bias will not be the same in each measure of the gap. This can be misleading and suggest erroneous interpretations regarding trends of the gaps over time (Reardon 2011). PISA has tried to make sure the tests are comparable across waves but it is still necessary to adjust for this imprecision (OECD 2012). Accordingly, each PISA survey provides a reliability indicator for each of the tests which can be used to adjust for the reliability of the scores.

In order to correct for this I calculate λ_i which is just the standardized test score $\hat{\gamma_i}$ adjusted by the reliability indicator of each wave. More formally, I calculate it through

$$\hat{\lambda_i} = \hat{\gamma_i} * \frac{1}{\sqrt{r}} \tag{2}$$

Where r is the reliability score of the test score in that PISA wave \footnote{Other procedures multiply each country by their own reliability measure for each year-subject pair (Chmielewski 2019). With this standardized and adjusted test score I define a dummy for those in the top 10% of the SES distribution and those at the bottom 10% of the SES distribution. I develop this standardization and estimation proceedure more in detail in the appendix. All models below are run separately for the bottom and top 10% of the SES distribution to test whether the autonomy measures have different dynamics for the two groups.

The model defined above is the standard model presented in the paper. However, I run a battery of different models to show the robustness of the results. Aside from the standard model, I run the same specification only on public schools, as the results of autonomy and test scores are certainly endogenous to the type of school. I also run a model for all available PISA countries (not only the developed subset of countries), a

[7] I also report results for literacy in the appendix

[8] Since each plausible value is a random draw from a theoretical latent normal distribution of possible student achievement scores, the median should be precise in getting a central measure of the latent distribution.

[9] This does not mess up the analysis by masking age-specific gaps as all students in the sample are 15 year olds. controlling for age is simply to adjust for monthly differences in ages.

[10] PISA 2000 has a slightly different metric over time

variant of the standard model that uses a linear model with country-year fixed effects (similarly to Hanushek, Link, and Woessmann (2013)), a model which exchanges the top and bottom 10% of students for the top and bottom 10% of schools, a model which runs all autonomy measures together (not separate models by autonomy measure, as was defined above) and a model with all students pooled and a formal interaction term between the SES groups and each autonomy measure. These models are presented in detail in the appendix and can be identified by their title. However, I also summarize their results in the results section.

2.3 Results

To begin, we visualize the highly contested SES achievement gap to understand the different dynamics between and within countries. Figure 2 shows the evolution of the achievement gap for developed countries.

As we can see from the results, some countries have increased their achievement strongly. For example, France, Austria and surprisingly Sweden have very steep slopes. France experienced an increase in inequality by roughly 0.9 SD, Austria by 0.6 and Sweden by 0.6. For such a short period of time, the magnitude of these increases are reasonably big.

Given that no one has estimated the evolution of the gap I cannot cross-check how other empirical estimations put France at. However, the work of Micklewright and Schnepf (2006) is the closest reference available which also finds that France was a low dispersion country in 2000; there is no evidence on what happened over time. Fortunately, the work of Bernardi and Ballarino (2016) did study social origin inequalities (broadly speaking, not in terms of achievement gaps) in France and found that they increased since the 2000's.

Other countries have reasonable increases such as Finland and Hungary, with increases of nearly 0.6 and 0.4 standard deviations respectively. Aside from these countries, there are other countries which experience no changes at all, specifically, Canada, Netherlands and Spain. Canada excels here not only because the gap has been stable over time, but because it has the smallest gap of all countries presented here. It is nearly 0.5 SD in 2000 and it increased only by 0.2 in 2015.

On the other hand, there are other countries which experience a decrease in the SES achievement gap. Poland decreased by about -0.4 and Denmark by -0.2. However, the most notable cases are the United States and Germany. These two countries show high levels of dispersion in the year 2000 with SES gaps of over 2 SD. But in the 15-year time trend both countries reduced the gaps by -0.6 and -0.9 respectively. Their distinctively large gaps in 2000 also show up in the work of Micklewright and Schnepf (2006). This finding is similar to the one in Reardon and Portilla (2016), in which they found a decreasing gap for kindergartners. It is important to highlight that the cohorts in their analysis are different from the ones in this study but also reassures that evidence close to the cohorts in this study also found a decline.

Analyzing figure 2 the reader may get the impression that these trends are not very steep and they should not be relevant in practical terms. However, note that the Y axis is measured in standard deviations. Small changes are actually large in practical terms. For example, evidence from PIRLS shows that the predicted growth of a student for a year of school is of around 0.30 standard deviations (Beaton and others 1996). PISA has also documented this type of metric in their annual reports (OECD. 2009). Take the case of Sweden. The slope does not look that steep but in reality it increased the gap from 1 SD in 2000 to around 1.5 SD in 2015. With that information in mind, the trends of Poland, United States, France and Germany gain particular relevance.

Table 2 contains the main models for the bottom 10% of all students and table 3 contains the main models for the top 10% of all students. Both tables have 2 models per autonomy measure corresponding to models for the index of academic autonomy, personnel autonomy and budget autonomy.

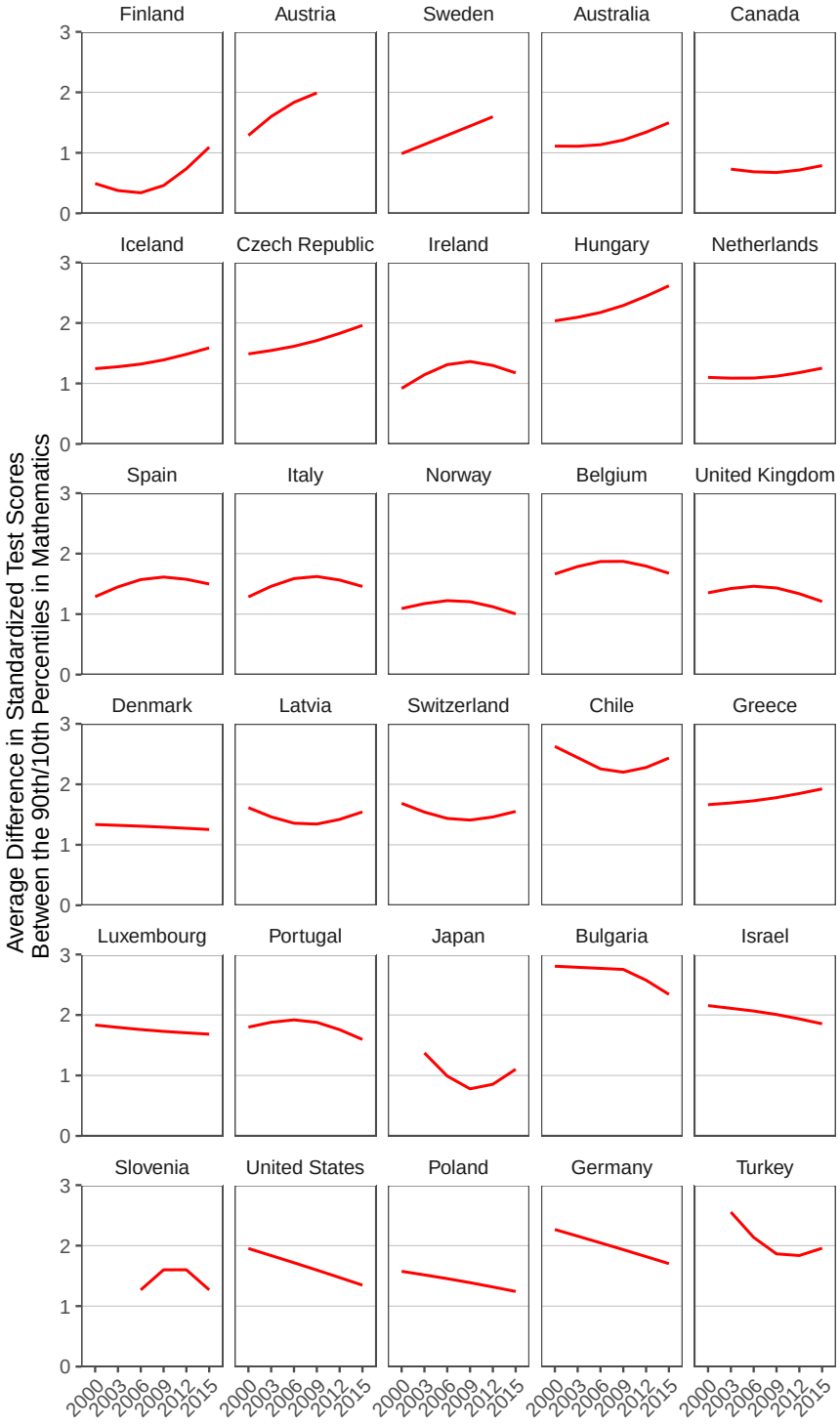

Figure 2: evolution of achievement gaps by countries

Table 2: Multilevel model with varying intercepts for bottom 10% of students in Mathematics

	Mathematics test score					
	Models restricted to bottom 10% of students					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	-0.053^{***} (0.012)	-0.065^{***} (0.012)				
Personnel autonomy			-0.036^{***} (0.013)	-0.077^{***} (0.013)		
Budget autonomy					−0.017 (0.011)	−0.011 (0.010)
- Gender: Male		0.193^{***} (0.007)		0.190^{***} (0.007)		0.193^{***} (0.007)
- Edu: Primary		−0.016 (0.014)		−0.017 (0.014)		−0.017 (0.014)
- Edu: Lower sec		0.027^{*} (0.014)		0.025^{*} (0.014)		0.023^{*} (0.014)
- Edu: Upper sec I		0.154^{***} (0.018)		0.150^{***} (0.018)		0.146^{***} (0.018)
- Edu: Upper sec II		0.063^{***} (0.016)		0.060^{***} (0.016)		0.054^{***} (0.016)
- Edu: University		-0.307^{***} (0.061)		-0.308^{***} (0.061)		-0.313^{***} (0.061)
- Books in HH: 11-100		0.278^{***} (0.008)		0.279^{***} (0.008)		0.277^{***} (0.008)
- Books in HH: 101-500		0.498^{***} (0.014)		0.502^{***} (0.014)		0.501^{***} (0.014)
- Books in HH: >500		0.178^{***} (0.035)		0.172^{***} (0.035)		0.177^{***} (0.035)
- Occupation index		0.005^{***} (0.0004)		0.005^{***} (0.0004)		0.005^{***} (0.0004)
- Native student		0.120^{***} (0.012)		0.118^{***} (0.012)		0.118^{***} (0.012)
- Voc track: Vocational		0.151^{***} (0.029)		0.145^{***} (0.029)		0.148^{***} (0.028)
- Voc track: General		0.236^{***} (0.026)		0.231^{***} (0.026)		0.231^{***} (0.026)
- Location: Town		-0.035^{**} (0.014)		-0.035^{**} (0.014)		-0.031^{**} (0.014)
- Location: Large town		-0.026^{*} (0.014)		-0.032^{**} (0.014)		-0.024^{*} (0.014)
- Location: City		-0.119^{***} (0.015)		-0.127^{***} (0.015)		-0.121^{***} (0.015)
- Location: Large city		-0.067^{***} (0.016)		-0.070^{***} (0.016)		-0.067^{***} (0.016)
- - Public (ref: private)		0.039^{**} (0.016)		0.007 (0.018)		0.044^{***} (0.016)
- Size of school		0.010^{***} (0.001)		0.010^{***} (0.001)		0.010^{***} (0.001)
- Constant	−0.104 (0.075)	-0.958^{***} (0.084)	−0.105 (0.075)	-0.906^{***} (0.083)	−0.107 (0.075)	-0.953^{***} (0.083)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.0041	0.004	0.0041	0.0039	0.0041	0.0039
Observations	52,488	52,488	52,641	52,641	52,732	52,732
Log Likelihood	−97,314.530	−95,566.860	−97,520.220	−95,767.070	−97,683.990	−95,937.410
Akaike Inf. Crit.	194,639.100	191,185.700	195,050.400	191,586.100	195,378.000	191,926.800
Bayesian Inf. Crit.	194,683.400	191,416.300	195,094.800	191,816.800	195,422.400	192,157.500

Note: *p<0.1; **p<0.05; ***p<0.01

Results for the first two models in table 2 shows that academic autonomy is associated with a decrease in test scores in Mathematics of about -0.06 standard deviations in test scores for students in the bottom 10%. These results hold even after controlling for the type of tracking programme that the child is enrolled in, the location of the school, the size of the school, the percentage of funding by the government, the percentage of certified teachers and whether the school is public or private. These results speak directly to the work of Stevenson and Baker (1991) which found that autonomy was detrimental in many cases but was criticized saying that these differences in autonomy was due to within country differences in school types. These results show that even after adjusting for different types of schools (private/public) and type of programme (different curricular tracks), the magnitude of the relationship is still big. These results seem to be as strong for personnel autonomy (model 4) with a decrease of nearly -0.07 points in the standardized mathematics test score on average for each country in every year.

Finally, the last two models for budget autonomy do not seem to be related to decreasing test scores, as the effect size is only -0.01 with a great deal of uncertainty. It is important to highlight that these effect size (-0.06 and -0.07) represent the average increase for the average year. If we multiply them to represent the actual time span of PISA (6 waves), it sums up to nearly 0.4 standard deviations. These results match very closely the results of Chmielewski (2019) in the evolution of the achievement gap and are quite big considering that magnitudes of over 0.3 standard deviations reflect a gap of about one grade's worth of knowledge. To put it simply, greater autonomy towards low performing students seems to be associated with a decrease of .4 standard deviations in test scores over the 15 years of data available.

Having said that, it can be the case that certain students self-select into schools with greater autonomy and that in itself is correlated with poorer performance. That is certainly playing a role in this estimation. However, it's not totally clear that it is the case given that within countries there is no particular evidence suggesting that public schools within a country have different levels of autonomy depending on performance.

Moving on to table 3, we can explore the results for the top 10% of students. The first two models show that academic autonomy seems to be positively associated with increasing student test scores by about .05 standard deviations for the top 10% of students. This amounts to a total of .3 standard deviations over the whole PISA waves. However, for the personnel and budget autonomy indexes both associations seem to be much weaker (effect sizes of about .01 for both) suggesting that academic autonomy seems to be an important autonomy component over both the good and bad performers yet autonomy over hiring teachers and setting their salaries is only detrimental in the context of bad performers.

Table 3: Multilevel model with varying intercepts for top 10% of students in Mathematics

	Mathematics test score					
	Models restricted to top 10% of students					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	0.072^{***} (0.010)	0.051^{***} (0.011)				
Personnel autonomy			0.035^{***} (0.010)	0.015 (0.011)		
Budget autonomy					0.034^{***} (0.010)	0.010 (0.010)
- Gender: Male		0.175^{***} (0.006)		0.178^{***} (0.006)		0.176^{***} (0.006)
- Edu: Primary		−0.102 (0.385)		−0.110 (0.386)		−0.099 (0.386)
- Edu: Lower sec		0.225 (0.373)		0.232 (0.374)		0.228 (0.373)
- Edu: Upper sec I		0.603 (0.372)		0.601 (0.373)		0.606 (0.373)
- Edu: Upper sec II		0.545^{***} (0.040)		0.541^{***} (0.041)		0.546^{***} (0.040)
- Edu: University		0.907^{***} (0.040)		0.906^{***} (0.040)		0.909^{***} (0.040)
- Books in HH: 11-100		1.025^{***} (0.040)		1.025^{***} (0.040)		1.029^{***} (0.040)
- Books in HH: 101-500		0.008^{***} (0.0003)		0.008^{***} (0.0003)		0.008^{***} (0.0003)
- Books in HH: >500		0.153^{***} (0.015)		0.155^{***} (0.015)		0.153^{***} (0.015)
- Occupation index		-0.312^{***} (0.033)		-0.304^{***} (0.033)		-0.310^{***} (0.033)
- Native student		0.278^{***} (0.029)		0.274^{***} (0.029)		0.274^{***} (0.029)
- Voc track: Vocational		−0.010 (0.017)		−0.006 (0.017)		−0.009 (0.017)
- Voc track: General		0.051^{***} (0.016)		0.058^{***} (0.016)		0.054^{***} (0.016)
- Location: Town		0.056^{***} (0.016)		0.066^{***} (0.016)		0.061^{***} (0.016)
- Location: Large town		0.060^{***} (0.018)		0.069^{***} (0.018)		0.065^{***} (0.018)
- Location: City		0.008 (0.015)		0.017 (0.015)		0.014 (0.015)
- Location: Large city		0.140^{***} (0.012)		0.145^{***} (0.013)		0.135^{***} (0.012)
- - Public (ref: private)		-0.002^{***} (0.0002)		-0.002^{***} (0.0002)		-0.002^{***} (0.0002)
- Size of school	1.172^{***} (0.074)	-1.449^{***} (0.396)	1.165^{***} (0.072)	-1.472^{***} (0.395)	1.170^{***} (0.074)	-1.455^{***} (0.395)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.0027	0.0046	0.0027	0.0044	0.0028	0.0045
Observations	72,713	72,713	72,683	72,683	72,767	72,767
Log Likelihood	−134,268.000	−131,065.000	−134,292.900	−131,099.500	−134,420.100	−131,214.500
Akaike Inf. Crit.	268,546.000	262,178.000	268,595.800	262,247.100	268,850.200	262,476.900
Bayesian Inf. Crit.	268,592.000	262,398.600	268,641.800	262,467.700	268,896.200	262,697.600

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4 and 5 replicate the results for literacy and show similar magnitudes and associations. In table 3 I find that academic autonomy is associated with a decrease in test scores of about -.09 standard deviations and personnel autonomy is associated with a decrease of -.04 standard deviations for the bottom 10% of students. Similarly, budget autonomy is unrelated to decreasing of increasing performance. However, in table 4 I find exactly the same results as for reading: for the top 10% of students, academic autonomy seems to be related to increasing test scores by about 0.05 standard deviations but not for personnel autonomy. However, budget autonomy seems to be associated with an increase in performance of about 0.02 standard deviations. The magnitude of this effect size is much lower than in the other models, so it is still relatively weak.

Table 4: Multilevel model with varying intercepts for bottom 10% of students in Literacy

	Reading test score					
	Models restricted to bottom 10% of students					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	-0.098^{***} (0.013)	-0.094^{***} (0.013)				
Personnel autonomy			-0.013 (0.014)	-0.040^{***} (0.015)		
Budget autonomy					-0.034^{***} (0.011)	-0.016 (0.011)
- Gender: Male		-0.276^{***} (0.008)		-0.280^{***} (0.008)		-0.279^{***} (0.008)
- Edu: Primary		0.107^{***} (0.016)		0.101^{***} (0.016)		0.101^{***} (0.016)
- Edu: Lower sec		0.107^{***} (0.015)		0.099^{***} (0.015)		0.099^{***} (0.015)
- Edu: Upper sec I		0.243^{***} (0.020)		0.233^{***} (0.020)		0.231^{***} (0.020)
- Edu: Upper sec II		0.118^{***} (0.018)		0.104^{***} (0.018)		0.103^{***} (0.018)
- Edu: University		-0.298^{***} (0.065)		-0.309^{***} (0.065)		-0.310^{***} (0.065)
- Books in HH: 11-100		0.318^{***} (0.008)		0.317^{***} (0.008)		0.317^{***} (0.008)
- Books in HH: 101-500		0.518^{***} (0.015)		0.520^{***} (0.015)		0.518^{***} (0.015)
- Books in HH: >500		0.175^{***} (0.039)		0.157^{***} (0.039)		0.169^{***} (0.039)
- Occupation index		0.006^{***} (0.0005)		0.006^{***} (0.0005)		0.006^{***} (0.0005)
- Native student		0.239^{***} (0.013)		0.237^{***} (0.013)		0.237^{***} (0.013)
- Voc track: Vocational		0.117^{***} (0.030)		0.099^{***} (0.030)		0.107^{***} (0.030)
- Voc track: General		0.257^{***} (0.028)		0.254^{***} (0.028)		0.252^{***} (0.028)
- Location: Town		0.016 (0.015)		0.018 (0.015)		0.021 (0.015)
- Location: Large town		0.059^{***} (0.015)		0.055^{***} (0.015)		0.057^{***} (0.015)
- Location: City		0.013 (0.016)		0.008 (0.016)		0.009 (0.016)
- Location: Large city		0.024 (0.018)		0.021 (0.018)		0.023 (0.018)
- - Public (ref: private)		0.029^{*} (0.017)		0.018 (0.019)		0.037^{**} (0.017)
- Size of school		0.013^{***} (0.001)		0.013^{***} (0.001)		0.013^{***} (0.001)
- Constant	-0.168^{**} (0.070)	-1.059^{***} (0.081)	-0.168^{**} (0.070)	-1.020^{***} (0.081)	-0.170^{**} (0.070)	-1.049^{***} (0.081)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.003	0.0029	0.0029	0.0028	0.0029	0.0028
Observations	52,256	52,256	52,410	52,410	52,504	52,504
Log Likelihood	−100,717.600	−98,389.350	−101,021.200	−98,662.810	−101,160.700	−98,813.020
Akaike Inf. Crit.	201,445.200	196,830.700	202,052.400	197,377.600	202,331.400	197,678.000
Bayesian Inf. Crit.	201,489.500	197,061.200	202,096.800	197,608.200	202,375.700	197,908.600

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Multilevel model with varying intercepts for top 10% of students in Literacy

	Reading test score					
	Models restricted to top 10% of students					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	0.079^{***} (0.011)	0.054^{***} (0.011)				
Personnel autonomy			0.034^{***} (0.010)	-0.013 (0.011)		
Budget autonomy					0.051^{***} (0.010)	0.022^{**} (0.009)
- Gender: Male		-0.259^{***} (0.006)		-0.259^{***} (0.006)		-0.258^{***} (0.006)
- Edu: Primary		0.146 (0.374)		0.140 (0.374)		0.152 (0.374)
- Edu: Lower sec		0.640^{*} (0.362)		0.641^{*} (0.362)		0.644^{*} (0.362)
- Edu: Upper sec I		0.938^{***} (0.361)		0.937^{***} (0.362)		0.942^{***} (0.361)
- Edu: Upper sec II		0.668^{***} (0.040)		0.672^{***} (0.040)		0.671^{***} (0.040)
- Edu: University		1.005^{***} (0.040)		1.011^{***} (0.040)		1.009^{***} (0.040)
- Books in HH: 11-100		1.123^{***} (0.040)		1.128^{***} (0.040)		1.127^{***} (0.040)
- Books in HH: 101-500		0.009^{***} (0.0003)		0.009^{***} (0.0003)		0.009^{***} (0.0003)
- Books in HH: >500		0.186^{***} (0.015)		0.186^{***} (0.015)		0.184^{***} (0.015)
- Occupation index		-0.240^{***} (0.032)		-0.234^{***} (0.032)		-0.239^{***} (0.032)
- Native student		0.298^{***} (0.028)		0.297^{***} (0.028)		0.294^{***} (0.028)
- Voc track: Vocational		-0.002 (0.017)		0.001 (0.017)		-0.005 (0.017)
- Voc track: General		0.080^{***} (0.016)		0.083^{***} (0.016)		0.078^{***} (0.016)
- Location: Town		0.110^{***} (0.017)		0.115^{***} (0.017)		0.112^{***} (0.017)
- Location: Large town		0.123^{***} (0.018)		0.127^{***} (0.018)		0.124^{***} (0.018)
- Location: City		0.053^{***} (0.015)		0.064^{***} (0.015)		0.057^{***} (0.015)
- Location: Large city		0.154^{***} (0.012)		0.143^{***} (0.013)		0.150^{***} (0.012)
- - Public (ref: private)		-0.002^{***} (0.0002)		-0.002^{***} (0.0002)		-0.002^{***} (0.0002)
- Size of school	1.116^{***} (0.067)	-1.955^{***} (0.383)	1.113^{***} (0.068)	-1.967^{***} (0.383)	1.115^{***} (0.068)	-1.957^{***} (0.383)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.0019	0.0037	0.0019	0.0036	0.0019	0.0037
Observations	72,306	72,306	72,277	72,277	72,361	72,361
Log Likelihood	−132,700.500	−128,850.600	−132,671.900	−128,821.100	−132,830.400	−128,998.500
Akaike Inf. Crit.	265,411.000	257,749.200	265,353.900	257,690.300	265,670.800	258,045.000
Bayesian Inf. Crit.	265,457.000	257,969.700	265,399.800	257,910.800	265,716.800	258,265.500

Note: *p<0.1; **p<0.05; ***p<0.01

As robustness, I've ran the previous model under several different specifications which I've compiled in figure 3.

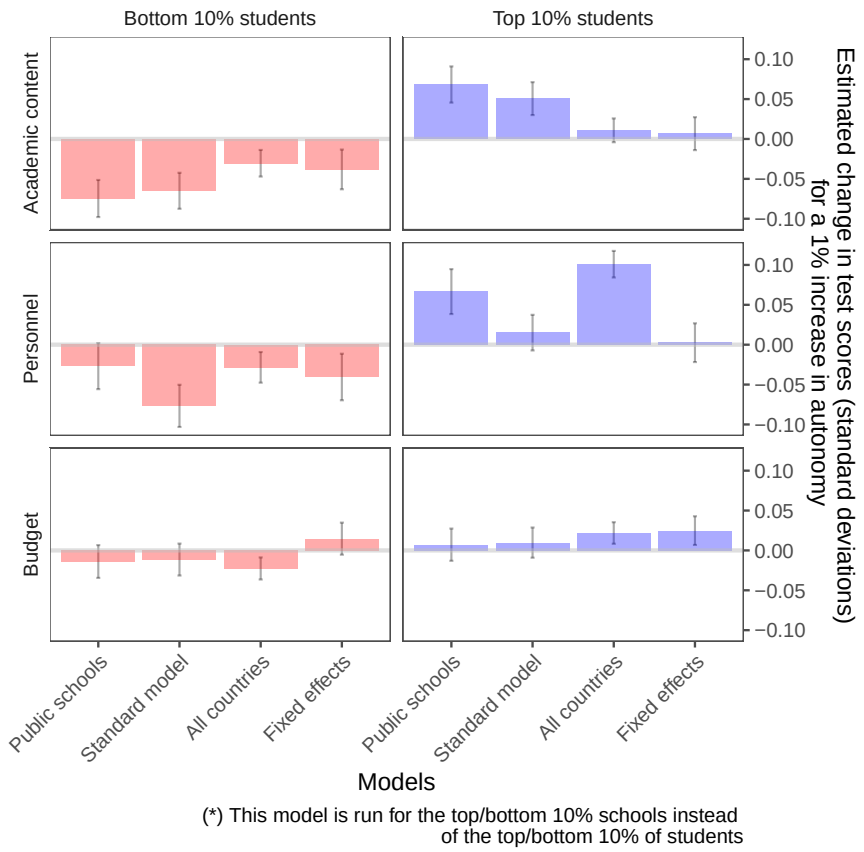

Figure 3: Comparison of autonomy measures under different model variations

On the right panel (red) we have all estimates related to the bottom 10% of students whereas on the left panel (blue) we have all estimates related to the top 10% of students. The x axis shows all different model specifications: "Public schools" refers to the same model but only restricted to public schools, "Standard model" refers to the previously shown models, "All countries" refers to models with all countries (including low income countries) and "Fixed effects" is an additional specification which uses a linear model with country-year fixed effects (to match Hanushek, Link, and Woessmann (2013)). The top left panel shows that the size of the effect sizes is very similar in all model specifications for academic autonomy. In contrast, the positive associations for the top 10% of students seem to dissapear once we add country-year fixed effects. In the middle panel, we see that for the personnel autonomy index, all models have negative relationships with the bottom 10% of students carrying sizable and robust associations. In contrast, for the top 10% of performers the relationship is quite sizable when restricted only to public schools and to all countries (increases of about 0.06-0.10 in standardized test scores) but dissapear when adding country-year fixed effects. Moving on to the bottom panel, results show that the effect sizes for the budget autonomy seem to be very small for both the top and bottom 10% of students. In summary, the evidence suggests that academic autonomy and personnel autonomy seem to have a somewhat sizable negative relationship with test scores yet this evidence is not symmetrical for the top 10% of students: the coefficients are somewhat sizable under certain specifications but the evidence is mixed. The models that estimated all of these changes are present in the appendix section. In addition, there are 4 tables concentrated not on the top/bottom 10% of students but on the top/bottom 10% of schools, to tackle directly whethere these results are similar when focusing only on the worst/best schools. Results are very similar.

3 Appendix

Table 6: Multilevel model with varying intercepts for bottom 10% of schools in Mathematics

	Mathematics test score					
	Models restricted to bottom 10% of schools					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	-0.042^{***} (0.015)	-0.029^{**} (0.015)				
Personnel autonomy			-0.006 (0.016)	-0.024 (0.017)		
Budget autonomy					-0.042^{***} (0.012)	-0.039^{***} (0.012)
- Gender: Male		0.117^{***} (0.008)		0.117^{***} (0.008)		0.117^{***} (0.008)
- Edu: Primary		0.060^{**} (0.029)		0.069^{**} (0.029)		0.069^{**} (0.029)
- Edu: Lower sec		0.097^{***} (0.027)		0.101^{***} (0.027)		0.103^{***} (0.027)
- Edu: Upper sec I		0.172^{***} (0.028)		0.174^{***} (0.028)		0.174^{***} (0.028)
- Edu: Upper sec II		0.144^{***} (0.027)		0.145^{***} (0.027)		0.147^{***} (0.027)
- Edu: University		0.054^{**} (0.027)		0.058^{**} (0.027)		0.059^{**} (0.027)
- Books in HH: 11-100		0.226^{***} (0.011)		0.227^{***} (0.011)		0.224^{***} (0.011)
- Books in HH: 101-500		0.526^{***} (0.013)		0.528^{***} (0.013)		0.526^{***} (0.013)
- Books in HH: >500		0.478^{***} (0.021)		0.473^{***} (0.021)		0.474^{***} (0.021)
- Occupation index		0.007^{***} (0.0003)		0.007^{***} (0.0003)		0.007^{***} (0.0003)
- Native student		0.138^{***} (0.014)		0.138^{***} (0.014)		0.138^{***} (0.014)
- Voc track: Vocational		0.209^{***} (0.025)		0.217^{***} (0.025)		0.211^{***} (0.025)
- Voc track: General		0.116^{***} (0.020)		0.116^{***} (0.020)		0.115^{***} (0.020)
- Location: Town		0.040^{***} (0.015)		0.038^{**} (0.015)		0.039^{***} (0.015)
- Location: Large town		0.002 (0.015)		-0.001 (0.015)		0.002 (0.015)
- Location: City		-0.008 (0.016)		-0.012 (0.016)		-0.008 (0.016)
- Location: Large city		-0.010 (0.021)		-0.018 (0.021)		-0.012 (0.021)
- - Public (ref: private)		0.003 (0.016)		-0.006 (0.017)		0.002 (0.016)
- Size of school		0.008^{***} (0.001)		0.009^{***} (0.001)		0.009^{***} (0.001)
- Constant	-0.549^{***} (0.097)	-1.519^{***} (0.096)	-0.550^{***} (0.096)	-1.513^{***} (0.096)	-0.549^{***} (0.096)	-1.521^{***} (0.096)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.2792	0.2587	0.2775	0.2596	0.2771	0.2581
Observations	34,519	34,519	34,600	34,600	34,631	34,631
Log Likelihood	$-41{,}331.840$	$-39{,}707.990$	$-41{,}454.670$	$-39{,}821.100$	$-41{,}492.070$	$-39{,}865.790$
Akaike Inf. Crit.	82,673.690	79,467.980	82,919.340	79,694.210	82,994.130	79,783.570
Bayesian Inf. Crit.	82,715.930	79,687.660	82,961.600	79,913.950	83,036.400	80,003.340

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: Multilevel model with varying intercepts for bottom 10% of schools in Literacy

	Reading test score					
	Models restricted to bottom 10% of schools					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	−0.001 (0.022)	0.005 (0.020)				
Personnel autonomy			0.067*** (0.022)	0.022 (0.023)		
Budget autonomy					0.015 (0.018)	−0.008 (0.017)
- Gender: Male		−0.375*** (0.012)		−0.373*** (0.012)		−0.376*** (0.012)
- Edu: Primary		0.171*** (0.040)		0.172*** (0.039)		0.176*** (0.040)
- Edu: Lower sec		0.169*** (0.037)		0.169*** (0.037)		0.172*** (0.037)
- Edu: Upper sec I		0.222*** (0.039)		0.218*** (0.039)		0.220*** (0.039)
- Edu: Upper sec II		0.241*** (0.037)		0.235*** (0.037)		0.239*** (0.037)
- Edu: University		0.131*** (0.038)		0.129*** (0.038)		0.133*** (0.038)
- Books in HH: 11-100		0.306*** (0.015)		0.310*** (0.015)		0.306*** (0.015)
- Books in HH: 101-500		0.615*** (0.019)		0.617*** (0.019)		0.614*** (0.019)
- Books in HH: >500		0.513*** (0.031)		0.512*** (0.031)		0.511*** (0.031)
- Occupation index		0.008*** (0.0004)		0.008*** (0.0004)		0.008*** (0.0004)
- Native student		0.147*** (0.020)		0.149*** (0.020)		0.150*** (0.020)
- Voc track: Vocational		0.038 (0.039)		0.047 (0.039)		0.033 (0.039)
- Voc track: General		0.093*** (0.029)		0.083*** (0.029)		0.085*** (0.029)
- Location: Town		0.033 (0.021)		0.030 (0.021)		0.032 (0.021)
- Location: Large town		0.017 (0.022)		0.014 (0.022)		0.016 (0.022)
- Location: City		0.008 (0.022)		−0.001 (0.022)		0.004 (0.022)
- Location: Large city		0.061** (0.028)		0.039 (0.028)		0.060** (0.028)
- - Public (ref: private)		−0.045** (0.022)		−0.029 (0.024)		−0.039* (0.022)
- Size of school		0.013*** (0.002)		0.013*** (0.002)		0.013*** (0.002)
- Constant	−0.705*** (0.110)	−1.525*** (0.109)	−0.716*** (0.110)	−1.525*** (0.110)	−0.707*** (0.110)	−1.522*** (0.109)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.2554	0.2094	0.2554	0.2138	0.2557	0.2104
Observations	22,325	22,325	22,353	22,353	22,402	22,402
Log Likelihood	−30,392.270	−28,927.170	−30,433.240	−28,980.000	−30,513.600	−29,048.760
Akaike Inf. Crit.	60,794.540	57,906.330	60,876.480	58,012.000	61,037.210	58,149.510
Bayesian Inf. Crit.	60,834.610	58,114.680	60,916.550	58,220.380	61,077.290	58,357.950

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Multilevel model with varying intercepts for top 10% of schools in Mathematics

	Mathematics test score					
	Models restricted to top 10% of schools					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	0.042^{***} (0.009)	0.030^{***} (0.009)				
Personnel autonomy			0.080^{***} (0.009)	0.042^{***} (0.011)		
Budget autonomy					0.010 (0.009)	−0.0004 (0.008)
- Gender: Male		0.228^{***} (0.005)		0.229^{***} (0.005)		0.229^{***} (0.005)
- Edu: Primary		0.113^{**} (0.057)		0.109^{*} (0.057)		0.106^{*} (0.057)
- Edu: Lower sec		0.174^{***} (0.053)		0.170^{***} (0.053)		0.167^{***} (0.053)
- Edu: Upper sec I		0.194^{***} (0.053)		0.191^{***} (0.053)		0.188^{***} (0.053)
- Edu: Upper sec II		0.149^{***} (0.052)		0.146^{***} (0.052)		0.143^{***} (0.052)
- Edu: University		0.195^{***} (0.052)		0.192^{***} (0.052)		0.189^{***} (0.052)
- Books in HH: 11-100		0.278^{***} (0.017)		0.276^{***} (0.017)		0.277^{***} (0.017)
- Books in HH: 101-500		0.506^{***} (0.017)		0.504^{***} (0.017)		0.505^{***} (0.017)
- Books in HH: >500		0.632^{***} (0.018)		0.629^{***} (0.018)		0.631^{***} (0.018)
- Occupation index		0.005^{***} (0.0002)		0.005^{***} (0.0002)		0.005^{***} (0.0002)
- Native student		0.047^{***} (0.011)		0.050^{***} (0.011)		0.047^{***} (0.011)
- Voc track: Vocational		0.064^{***} (0.024)		0.058^{**} (0.024)		0.062^{***} (0.024)
- Voc track: General		0.013 (0.017)		0.011 (0.017)		0.015 (0.017)
- Location: Town		0.029^{*} (0.017)		0.029^{*} (0.017)		0.028 (0.017)
- Location: Large town		0.023 (0.017)		0.023 (0.017)		0.021 (0.017)
- Location: City		0.009 (0.017)		0.008 (0.017)		0.006 (0.017)
- Location: Large city		0.059^{***} (0.018)		0.058^{***} (0.018)		0.058^{***} (0.018)
- - Public (ref: private)		0.043^{***} (0.009)		0.055^{***} (0.010)		0.038^{***} (0.009)
- Size of school		0.001 (0.001)		0.001 (0.001)		0.001 (0.001)
- Constant	1.434^{***} (0.077)	0.323^{***} (0.097)	1.427^{***} (0.078)	0.313^{***} (0.097)	1.432^{***} (0.078)	0.329^{***} (0.097)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.1987	0.2155	0.2023	0.2181	0.2018	0.2177
Observations	68,227	68,227	68,319	68,319	68,328	68,328
Log Likelihood	−73,692.400	−70,252.040	−73,769.970	−70,332.390	−73,816.380	−70,341.730
Akaike Inf. Crit.	147,394.800	140,556.100	147,549.900	140,716.800	147,642.800	140,735.400
Bayesian Inf. Crit.	147,440.500	140,793.500	147,595.600	140,954.200	147,688.400	140,972.900

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9: Multilevel model with varying intercepts for top 10% of schools in Literacy

	Reading test score					
	Models restricted to top 10% of schools					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	0.043*** (0.009)	0.034*** (0.008)				
Personnel autonomy			0.071*** (0.009)	0.042*** (0.010)		
Budget autonomy					0.013 (0.008)	0.015* (0.008)
- Gender: Male		−0.245*** (0.005)		−0.244*** (0.005)		−0.244*** (0.005)
- Edu: Primary		0.088 (0.058)		0.095 (0.058)		0.088 (0.058)
- Edu: Lower sec		0.105* (0.054)		0.112** (0.055)		0.107* (0.055)
- Edu: Upper sec I		0.183*** (0.054)		0.187*** (0.055)		0.183*** (0.054)
- Edu: Upper sec II		0.147*** (0.054)		0.152*** (0.054)		0.146*** (0.054)
- Edu: University		0.179*** (0.054)		0.184*** (0.054)		0.178*** (0.054)
- Books in HH: 11-100		0.308*** (0.015)		0.307*** (0.015)		0.308*** (0.015)
- Books in HH: 101-500		0.532*** (0.015)		0.532*** (0.015)		0.533*** (0.015)
- Books in HH: >500		0.641*** (0.016)		0.642*** (0.016)		0.642*** (0.016)
- Occupation index		0.005*** (0.0002)		0.005*** (0.0002)		0.005*** (0.0002)
- Native student		0.099*** (0.010)		0.101*** (0.010)		0.100*** (0.010)
- Voc track: Vocational		−0.116*** (0.021)		−0.115*** (0.021)		−0.111*** (0.021)
- Voc track: General		0.079*** (0.016)		0.079*** (0.016)		0.082*** (0.016)
- Location: Town		0.064*** (0.014)		0.062*** (0.014)		0.061*** (0.014)
- Location: Large town		0.106*** (0.014)		0.104*** (0.014)		0.102*** (0.014)
- Location: City		0.132*** (0.014)		0.130*** (0.014)		0.127*** (0.014)
- Location: Large city		0.176*** (0.015)		0.173*** (0.015)		0.173*** (0.015)
- - Public (ref: private)		0.023*** (0.008)		0.034*** (0.009)		0.019** (0.008)
- Size of school		0.001 (0.001)		0.001 (0.001)		0.001 (0.001)
- Constant	1.285*** (0.051)	0.202** (0.079)	1.280*** (0.050)	0.188** (0.079)	1.283*** (0.050)	0.203*** (0.079)
N. Country	29	29	29	29	29	29
N. Wave	6	6	6	6	6	6
ICC	0.0957	0.1014	0.0973	0.1038	0.097	0.1027
Observations	81,766	81,766	81,732	81,732	81,803	81,803
Log Likelihood	−88,849.870	−84,017.570	−88,821.870	−84,011.170	−88,920.000	−84,084.860
Akaike Inf. Crit.	177,709.700	168,087.100	177,653.700	168,074.300	177,850.000	168,221.700
Bayesian Inf. Crit.	177,756.300	168,329.200	177,700.300	168,316.400	177,896.600	168,463.800

Note: *p<0.1; **p<0.05; ***p<0.01

Table 10: Multilevel model with varying intercepts for bottom 10% of students in Mathematics for all countries (not only developed)

	Mathematics test score					
	Models restricted to bottom 10% of students					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	-0.029^{***} (0.009)	-0.030^{***} (0.008)				
Personnel autonomy			-0.038^{***} (0.008)	-0.028^{***} (0.010)		
Budget autonomy					-0.021^{***} (0.007)	-0.023^{***} (0.007)
- Gender: Male		0.158^{***} (0.005)		0.158^{***} (0.005)		0.158^{***} (0.005)
- Edu: Primary		0.019^{**} (0.008)		0.017^{**} (0.007)		0.018^{**} (0.007)
- Edu: Lower sec		0.006 (0.010)		0.004 (0.010)		0.005 (0.010)
- Edu: Upper sec I		0.123^{***} (0.015)		0.120^{***} (0.015)		0.119^{***} (0.015)
- Edu: Upper sec II		0.066^{***} (0.013)		0.062^{***} (0.013)		0.062^{***} (0.013)
- Edu: University		-0.314^{***} (0.056)		-0.315^{***} (0.056)		-0.315^{***} (0.056)
- Books in HH: 11-100		0.171^{***} (0.006)		0.170^{***} (0.006)		0.170^{***} (0.006)
- Books in HH: 101-500		0.360^{***} (0.012)		0.362^{***} (0.012)		0.362^{***} (0.012)
- Books in HH: >500		0.122^{***} (0.030)		0.118^{***} (0.030)		0.120^{***} (0.030)
- Occupation index		0.004^{***} (0.0003)		0.004^{***} (0.0003)		0.004^{***} (0.0003)
- Native student		0.163^{***} (0.011)		0.162^{***} (0.011)		0.161^{***} (0.011)
- Voc track: Vocational		0.259^{***} (0.022)		0.261^{***} (0.022)		0.261^{***} (0.022)
- Voc track: General		0.282^{***} (0.020)		0.285^{***} (0.020)		0.282^{***} (0.020)
- Location: Town		0.038^{***} (0.008)		0.039^{***} (0.008)		0.042^{***} (0.008)
- Location: Large town		0.071^{***} (0.009)		0.070^{***} (0.009)		0.073^{***} (0.009)
- Location: City		0.027^{***} (0.010)		0.024^{**} (0.010)		0.025^{**} (0.010)
- Location: Large city		0.034^{***} (0.013)		0.031^{**} (0.013)		0.032^{**} (0.013)
- - Public (ref: private)		0.118^{***} (0.010)		0.106^{***} (0.012)		0.120^{***} (0.010)
- Size of school		0.009^{***} (0.0004)		0.009^{***} (0.0004)		0.009^{***} (0.0004)
- Constant	-0.377^{***} (0.076)	-1.250^{***} (0.080)	-0.378^{***} (0.076)	-1.231^{***} (0.080)	-0.377^{***} (0.076)	-1.246^{***} (0.080)
N. Country	69	69	69	69	69	69
N. Wave	6	6	6	6	6	6
ICC	0.0096	0.0096	0.0096	0.0096	0.0095	0.0096
Observations	90,304	90,304	90,452	90,452	90,552	90,552
Log Likelihood	−165,896.300	−163,915.500	−166,075.500	−164,095.100	−166,239.700	−164,251.700
Akaike Inf. Crit.	331,802.500	327,882.900	332,161.000	328,242.200	332,489.500	328,555.400
Bayesian Inf. Crit.	331,849.600	328,127.600	332,208.100	328,486.900	332,536.600	328,800.200

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: Country-wave fixed effect models for bottom 10% of students in Mathematics

	Mathematics test score					
	Models restricted to bottom 10% of students					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	0.085^{***} (0.011)	-0.038^{***} (0.013)				
Personnel autonomy			0.109^{***} (0.011)	-0.040^{***} (0.015)		
Budget autonomy					0.055^{***} (0.011)	0.015 (0.010)
- Gender: Male		0.185^{***} (0.007)		0.185^{***} (0.007)		0.185^{***} (0.007)
- Edu: Primary		0.123^{***} (0.016)		0.124^{***} (0.015)		0.123^{***} (0.015)
- Edu: Lower sec		0.163^{***} (0.016)		0.164^{***} (0.016)		0.163^{***} (0.016)
- Edu: Upper sec I		0.253^{***} (0.019)		0.253^{***} (0.019)		0.250^{***} (0.019)
- Edu: Upper sec II		0.189^{***} (0.019)		0.190^{***} (0.018)		0.188^{***} (0.018)
- Edu: University		-0.246^{***} (0.052)		-0.243^{***} (0.052)		-0.245^{***} (0.052)
- Books in HH: 11-100		0.287^{***} (0.008)		0.287^{***} (0.008)		0.287^{***} (0.008)
- Books in HH: 101-500		0.547^{***} (0.013)		0.547^{***} (0.013)		0.548^{***} (0.013)
- Books in HH: >500		0.343^{***} (0.033)		0.338^{***} (0.033)		0.343^{***} (0.033)
- Occupation index		0.003^{***} (0.0005)		0.003^{***} (0.0005)		0.003^{***} (0.0005)
- Native student		0.229^{***} (0.013)		0.226^{***} (0.013)		0.226^{***} (0.013)
- Voc track: Vocational		0.004 (0.021)		0.009 (0.021)		0.005 (0.021)
- Voc track: General		0.211^{***} (0.018)		0.210^{***} (0.018)		0.208^{***} (0.018)
- Location: Town		0.016 (0.013)		0.018 (0.013)		0.017 (0.013)
- Location: Large town		-0.003 (0.013)		-0.003 (0.013)		-0.003 (0.013)
- Location: City		-0.031^{**} (0.015)		-0.033^{**} (0.015)		-0.034^{**} (0.015)
- Location: Large city		-0.011 (0.020)		-0.006 (0.020)		-0.011 (0.020)
- - Public (ref: private)		-0.091^{***} (0.015)		-0.104^{***} (0.016)		-0.086^{***} (0.015)
- Size of school		0.020^{***} (0.001)		0.020^{***} (0.001)		0.020^{***} (0.001)
- Constant	-0.052^{***} (0.004)	-0.945^{***} (0.045)	-0.052^{***} (0.004)	-0.956^{***} (0.045)	-0.054^{***} (0.004)	-0.965^{***} (0.045)
Country-Year fixed effects	yes	yes	yes	yes	yes	yes
Observations	52,488	52,488	52,641	52,641	52,732	52,732
R^2	0.001	0.172	0.002	0.171	0.001	0.171
Adjusted R^2	0.001	0.171	0.002	0.170	0.0005	0.171
Residual Std. Error	0.908 (df = 52486)	0.828 (df = 52432)	0.908 (df = 52639)	0.827 (df = 52585)	0.909 (df = 52730)	0.828 (df = 52676)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 12: Country-wave fixed effect models for top 10% of students in Mathematics

	Mathematics test score					
	Models restricted to top 10% of students					
	(1)	(2)	(3)	(4)	(5)	(6)
Academic autonomy	0.123^{***} (0.009)	0.007 (0.011)				
Personnel autonomy			0.141^{***} (0.008)	0.003 (0.012)		
Budget autonomy					0.028^{***} (0.010)	0.025^{***} (0.009)
- Gender: Male		0.172^{***} (0.006)		0.172^{***} (0.006)		0.173^{***} (0.006)
- Edu: Primary		0.186 (0.322)		0.186 (0.322)		0.207 (0.322)
- Edu: Lower sec		0.412 (0.303)		0.414 (0.303)		0.414 (0.303)
- Edu: Upper sec I		0.581^{*} (0.303)		0.582^{*} (0.302)		0.582^{*} (0.302)
- Edu: Upper sec II		0.613^{***} (0.040)		0.609^{***} (0.040)		0.614^{***} (0.039)
- Edu: University		0.957^{***} (0.039)		0.953^{***} (0.039)		0.958^{***} (0.039)
- Books in HH: 11-100		1.096^{***} (0.039)		1.091^{***} (0.039)		1.096^{***} (0.039)
- Books in HH: 101-500		0.008^{***} (0.0003)		0.008^{***} (0.0003)		0.008^{***} (0.0003)
- Books in HH: >500		0.140^{***} (0.012)		0.141^{***} (0.012)		0.140^{***} (0.012)
- Occupation index		-0.387^{***} (0.024)		-0.387^{***} (0.024)		-0.390^{***} (0.024)
- Native student		0.195^{***} (0.018)		0.196^{***} (0.018)		0.194^{***} (0.018)
- Voc track: Vocational		0.022 (0.015)		0.023 (0.015)		0.019 (0.015)
- Voc track: General		0.067^{***} (0.015)		0.069^{***} (0.015)		0.065^{***} (0.015)
- Location: Town		0.091^{***} (0.015)		0.094^{***} (0.015)		0.091^{***} (0.015)
- Location: Large town		0.136^{***} (0.017)		0.136^{***} (0.017)		0.134^{***} (0.017)
- Location: City		0.020 (0.014)		0.019 (0.014)		0.022 (0.014)
- Location: Large city		-0.027^{***} (0.010)		-0.027^{**} (0.012)		-0.026^{**} (0.010)
- - Public (ref: private)		-0.002^{***} (0.0002)		-0.002^{***} (0.0002)		-0.002^{***} (0.0002)
- Size of school	1.189^{***} (0.003)	-1.370^{***} (0.309)	1.188^{***} (0.003)	-1.387^{***} (0.309)	1.187^{***} (0.003)	-1.388^{***} (0.308)
Country-Year fixed effects	yes	yes	yes	yes	yes	yes
Observations	72,713	72,713	72,683	72,683	72,767	72,767
R^2	0.003	0.179	0.004	0.179	0.0001	0.179
Adjusted R^2	0.003	0.178	0.004	0.178	0.0001	0.178
Residual Std. Error	0.881 (df = 72711)	0.800 (df = 72659)	0.881 (df = 72681)	0.800 (df = 72629)	0.882 (df = 72765)	0.800 (df = 72713)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13: Multilevel model with varying intercepts for bottom 10% of students in Mathematics with all autonomy measures pooled

	Mathematics test score	
	Models restricted to bottom 10% of students	
	(1)	(2)
Academic autonomy	-0.044^{***} (0.013)	-0.048^{***} (0.012)
Personnel autonomy	-0.018 (0.014)	-0.059^{***} (0.015)
Budget autonomy	-0.006 (0.011)	0.006 (0.011)
- Gender: Male		0.192^{***} (0.007)
- Edu: Primary		-0.016 (0.014)
- Edu: Lower sec		0.029^{**} (0.014)
- Edu: Upper sec I		0.157^{***} (0.018)
- Edu: Upper sec II		0.066^{***} (0.016)
- Edu: University		-0.303^{***} (0.061)
- Books in HH: 11-100		0.279^{***} (0.008)
- Books in HH: 101-500		0.502^{***} (0.014)
- Books in HH: >500		0.180^{***} (0.035)
- Occupation index		0.005^{***} (0.0004)
- Native student		0.119^{***} (0.012)
- Voc track: Vocational		0.153^{***} (0.029)
- Voc track: General		0.236^{***} (0.026)
- Location: Town		-0.033^{**} (0.014)
- Location: Large town		-0.027^{*} (0.014)
- Location: City		-0.122^{***} (0.015)
- Location: Large city		-0.061^{***} (0.017)
- - Public (ref: private)		0.011 (0.018)
- Size of school		0.010^{***} (0.001)
- Constant	-0.104 (0.075)	-0.899^{***} (0.084)
N. Country	29	29
N. Wave	6	6
ICC	0.0041	0.0039
Observations	52,189	52,189
Log Likelihood	$-96{,}817.440$	$-95{,}064.800$
Akaike Inf. Crit.	193,648.900	190,185.600
Bayesian Inf. Crit.	193,710.900	190,433.800
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 14: Multilevel model with varying intercepts for top 10% of students in Mathematics with all autonomy measures pooled

	Mathematics test score	
	Models restricted to top 10% of students	
	(1)	(2)
Academic autonomy	0.066^{***} (0.012)	0.052^{***} (0.011)
Personnel autonomy	0.003 (0.011)	−0.008 (0.012)
Budget autonomy	0.014 (0.011)	−0.003 (0.010)
- Gender: Male		0.176^{***} (0.006)
- Edu: Primary		−0.098 (0.386)
- Edu: Lower sec		0.234 (0.373)
- Edu: Upper sec I		0.601 (0.373)
- Edu: Upper sec II		0.548^{***} (0.041)
- Edu: University		0.912^{***} (0.040)
- Books in HH: 11-100		1.029^{***} (0.040)
- Books in HH: 101-500		0.008^{***} (0.0003)
- Books in HH: >500		0.153^{***} (0.015)
- Occupation index		-0.310^{***} (0.033)
- Native student		0.282^{***} (0.029)
- Voc track: Vocational		−0.009 (0.017)
- Voc track: General		0.053^{***} (0.016)
- Location: Town		0.058^{***} (0.016)
- Location: Large town		0.064^{***} (0.018)
- Location: City		0.008 (0.015)
- Location: Large city		0.136^{***} (0.013)
- - Public (ref: private)		-0.002^{***} (0.0002)
- Size of school	1.172^{***} (0.074)	-1.467^{***} (0.396)
N. Country	29	29
N. Wave	6	6
ICC	0.0027	0.0046
Observations	72,244	72,244
Log Likelihood	−133,507.400	−130,317.100
Akaike Inf. Crit.	267,028.800	260,686.300
Bayesian Inf. Crit.	267,093.100	260,925.200
Note:		*p<0.1; **p<0.05; ***p<0.01

3.1 Calculating Achievement Gaps

To *standardize* the test score I fit a linear model

$$y_i = \alpha + \beta_1 * AGE_i + \epsilon_i, \quad \epsilon_i \ N(0, \sigma^2) \tag{3}$$

for each wave, where y_i is the median student test score for student i and AGE_i is their age measured in months (following the same strategy as Reardon (2011) [11]) weighted by the student sample weights from PISA [12].

I then calculate $\hat{\gamma_i}$ by

$$\hat{\gamma_i} = \frac{\hat{\epsilon_i}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}} \tag{4}$$

where $\hat{\epsilon_i}$ is the residual for student i, $\hat{y_i}$ is the predicted test score for student i and the denominator is the root mean square error of the model.

This new standardized variable has a mean of zero. Standardizing the median test score solves the problem of comparability between different tests and across waves as the test scores have now the same metric across time. Another concern is whether test scores measured at different waves have different amounts of measurement error. If that is the case, then the amount of bias will not be the same in each measure of the gap. This can be misleading and suggest erroneous interpretations regarding trends of the gaps over time (Reardon 2011). PISA has tried to make sure the tests are comparable across waves but it is still necessary to adjust for this imprecision (OECD 2012). Accordingly, each PISA survey provides a reliability indicator for each of the tests which can be used to adjust for the reliability of the scores.

In order to correct for this I calculate λ_i which is just $\hat{\gamma_i}$ adjusted by the reliability indicator of each wave. More formally, I calculate it through

$$\hat{\lambda_i} = \hat{\gamma_i} * \frac{1}{\sqrt{r}} \tag{5}$$

where r is the reliability score of the test score in that PISA wave.[13] Note that I implement equation (5) separately by test scores and waves because there is a separate reliability indicator for each one. Once that is adjusted, the test scores should be roughly free of any bias in the trend that may arise from differential reliability of the tests.

In order to calculate the SES gaps it is necessary to estimate the thresholds for the 90th and 10th percentile. I calculate the thresholds using the SES index separately for each country-wave combination using the specific student sample weights of each one. I then generate a dummy of 1 for those above (including) the 90th percentile and 0 for those below (including) the 10th percentile for each country-wave pair.

I then fit a multilevel model:

$$\lambda_{ij} = \alpha_j + \beta_j * SES_i + \epsilon_{ij}, \text{ for } i = 1, 2, ..., n \text{ for each country } j \tag{6}$$

[11] This does not mess up the analysis by masking age-specific gaps as all students in the sample are 15 year olds. Controlling for age is simply to adjust for monthly differences in ages.

[12] I also tried to run the model for each country-wave separately but the results were very similar and it was more computationally expensive

[13] Other procedures multiply each country by their own reliability measure for each year-subject pair (Chmielewski 2019). The reliability estimates are calculated using Item Response Theory (IRT) analogues of traditional estimates of person separation reliability such as internal consistency. Unfortunately, PISA 2000 did not provide any reliability measure separately for each country and at the moment of the writing of this paper, PISA 2015 has yet to release their own. For these reasons, I implement the analysis following the original work of (Reardon 2011)

where SES_i is whether the student is at or above the 90th percentile (coded as 1) or whether it is at or below the 10th percentile (coded as 0). I allow α and β to vary by country j in order to obtain gaps for each country. I implement this model separately for each wave and weight by the wave-specific student sample weights. The previous model allows to calculate the achievement gap for each country by extracting the β's and α's for each country. I also calculate the standard error of this difference and generate uncertainty intervals.

Bibliography

Alvaredo, Facundo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2017. "Global Inequality Dynamics: New Findings from Wid. World." *American Economic Review* 107 (5): 404–09.

Ammermüller, Andreas. 2005. "Educational Opportunities and the Role of Institutions." *ZEW-Centre for European Economic Research Discussion Paper*, nos. 05-044.

Beaton, Albert E, and others. 1996. *Mathematics Achievement in the Middle School Years. IEA's Third International Mathematics and Science Study (Timss).* ERIC.

Bernardi, Fabrizio, and Gabriele Ballarino. 2016. *Education, Occupation and Social Origin: A Comparative Analysis of the Transmission of Socio-Economic Inequalities.* Edward Elgar Publishing.

Bradbury, Bruce, Miles Corak, Jane Waldfogel, and Elizabeth Washbrook. 2015. *Too Many Children Left Behind: The Us Achievement Gap in Comparative Perspective.* Russell Sage Foundation.

Broer, Markus, Yifan Bai, and Frank Fonseca. 2019. *Socioeconomic Inequality and Educational Outcomes: Evidence from Twenty Years of Timss.* Vol. 5. Springer.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos. 2011. *Making Schools Work: New Evidence on Accountability Reforms.* World Bank Publishing. https://www.ebook.de/de/product/14536372/barbara_bruns_deon_filmer_harry_anthony_patrinos_making_schools_work_new_evidence_on_accountability_reforms.html.

Campbell, Frances A, Craig T Ramey, Elizabeth Pungello, Joseph Sparling, and Shari Miller-Johnson. 2002. "Early Childhood Education: Young Adult Outcomes from the Abecedarian Project." *Applied Developmental Science* 6 (1): 42–57.

Chmielewski, Anna K. 2019. "The Global Increase in the Socioeconomic Achievement Gap, 1964 to 2015." *American Sociological Review*, 0003122419847165.

Chmielewski, Anna K, and Sean F Reardon. 2016. "Patterns of Cross-National Variation in the Association Between Income and Academic Achievement." *AERA Open* 2 (3): 2332858416649593.

Cunha, Flavio, James J Heckman, Lance Lochner, and Dimitriy V Masterov. 2006. "Interpreting the Evidence on Life Cycle Skill Formation." *Handbook of the Economics of Education* 1: 697–812.

Di Gropello, Emanuela. 2006. *A Comparative Analysis of School-Based Management in Central America.* 72. World Bank Publications.

Dupriez, Vincent, and Xavier Dumay. 2006. "Inequalities in School Systems: Effect of School Structure or of Society Structure?" *Comparative Education* 42 (02): 243–60.

Duru-Bellat, Marie, and Bruno Suchaut. 2005. "Organisation and Context, Efficiency and Equity of Educational Systems: What Pisa Tells Us." *European Educational Research Journal* 4 (3): 181–94.

Gamoran, Adam, and Mark Berends. 1987. "The Effects of Stratification in Secondary Schools: Synthesis of Survey and Ethnographic Research." *Review of Educational Research* 57 (4): 415–35.

Hanushek, Eric A, Susanne Link, and Ludger Woessmann. 2013. "Does School Autonomy Make Sense Everywhere? Panel Estimates from Pisa." *Journal of Development Economics* 104: 212–32.

Hanushek, Eric A, Paul E Peterson, Laura M Talpey, and Ludger Woessmann. 2019. "The Unwavering Ses Achievement Gap: Trends in Us Student Performance." National Bureau of Economic Research.

Hanushek, Eric, and others. 2006. "Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries." *The Economic Journal* 116 (510): C63–C76.

Hattie, John AC. 2002. "Classroom Composition and Peer Effects." *International Journal of Educational Research* 37 (5): 449–81.

Kulic, Nevena, Jan Skopek, Moris Triventi, and Hans-Peter Blossfeld. 2019. "Social Background and Children's Cognitive Skills: The Role of Early Childhood Education and Care in a Cross-National Perspective." *Annual Review of Sociology* 45.

LeTendre, Gerald K, Barbara K Hofer, and Hidetada Shimizu. 2003. "What Is Tracking? Cultural Expectations in the United States, Germany, and Japan." *American Educational Research Journal* 40 (1): 43–89.

Magnuson, Katherine, and Jane Waldfogel. 2008. *Steady Gains and Stalled Progress: Inequality and the Black-White Test Score Gap*. Russell Sage Foundation.

Marks, Gary N, John Cresswell, and John Ainley. 2006. "Explaining Socioeconomic Inequalities in Student Achievement: The Role of Home and School Factors." *Educational Research and Evaluation* 12 (02): 105–28.

Micklewright, John, and Sylke V Schnepf. 2006. "Inequality of Learning in Industrialised Countries."

Milanovic, Branko. 2016. *Global Inequality: A New Approach for the Age of Globalization*. Harvard University Press.

OECD. 2002. *Education at a Glance: OECD Indicators 2002*. OECD Publishing.

———. 2012. *PISA 2009 Technical Report*. OECD Publishing.

———. 2016. *PISA 2015 Results (Volume I)*. OECD Publishing.

OECD. 2009. "PISA Data Analysis Manual: SAS 2nd Edition." OECD Paris, France.

Reardon, Sean F. 2011. "The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations." *Whither Opportunity*, 91–116.

Reardon, Sean F, and Ximena A Portilla. 2016. "Recent Trends in Income, Racial, and Ethnic School Readiness Gaps at Kindergarten Entry." *Aera Open* 2 (3): 2332858416657343.

Reardon, Sean F, Joseph P Robinson, and Ericka S Weathers. 2008. "Patterns and Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps." *Handbook of Research in Education Finance and Policy*, 497–516.

Stevenson, David Lee, and David P Baker. 1991. "State Control of the Curriculum and Classroom Instruction." *Sociology of Education*, 1–10.

Vandenberge, Vincent. 2006. "Achievement Effectiveness and Equity: The Role of Tracking, Grade Repetition and Inter-School Segregation." *Applied Economics Letters* 13 (11): 685–93.

Van de Werfhorst, Herman G, and Jonathan JB Mijs. 2010. "Achievement Inequality and the Institutional Structure of Educational Systems: A Comparative Perspective." *Annual Review of Sociology* 36: 407–28.

Wu, Margaret. 2005. "The Role of Plausible Values in Large-Scale Surveys." *Studies in Educational Evaluation* 31 (2-3): 114–28.