Does the Impact of Ability Grouping vary with the Culture of Competitiveness? - Evidence from PISA 2012 *

Kathrin Thiemann†

March 15, 2017

Abstract

In this paper theoretical hypotheses from Thiemann (2016) are tested for their empirical relevance. According to theory comprehensive schooling and ability grouping yield different results in terms of average student performance in countries that differ in their culture of competitiveness. The predictions are tested using a country-level indicator on the appraisal of competition from the World Values Survey. Educational achievement data is from PISA 2012, covering 34 countries and more than 10,000 schools of which data on the school's policy of ability grouping is available. To overcome possible endogeneity of ability grouping an instrumental variable approach is employed, using the number of schools a school regionally competes with as an instrument. The estimation shows that ability grouping in some or all classes increases average student achievement in competitive cultures and decreases average student achievement in non-competitive cultures.

JEL-Code: I20, I24, O15, H75

Keywords: Ability Grouping, Ability Tracking, Culture, Competitiveness, PISA, Education Production Function, Instrumental Variables, Quantile Regression

*I would like to thank my second supervisor Prof. Thomas Siedler (PhD) for helpful comments.

†Department of Economics, University of Hamburg, Von-Melle-Park 5, D-20146-Hamburg, Germany. Phone: +49 40 42838 6331. Email: kathrin.thiemann@uni-hamburg.de. Web: www.wiso.uni-hamburg.de/io

1 Introduction

Among the top performers of the most recent PISA (Programme for International Student Assessment) study 2012 are countries like Switzerland, the Netherlands or Singapore (OECD, 2013b), all countries that rigidly sort students into different schools based on their ability. Still, there are also countries like Finland and Japan at the top of the ranking, where students of very heterogeneous abilities are all taught together in one class. This suggests that different approaches are successful in different countries. In a recently published report "The learning curve" (The Economist Intelligence Unit, 2012) about the search for international best practices in education, the authors admit that none were found. They describe the way in which differences in the country-specific learning process transform inputs into outputs as a "black box" which is difficult to predict or quantify consistently. A possible reason for this finding is that countries differ in their cultures of teaching and learning. The question focused on in this paper is whether learning in small ability segregated groups (ability grouping) is to be preferred over learning in a class with students of heterogeneous abilities and backgrounds (comprehensive schooling) and to what degree the answer depends on student characteristics that vary with culture. In particular we focus on the country-specific culture of competitiveness that might influence the effect of AG on student achievement. Competitiveness thereby refers to the innate drive and desire of students to socially compare and outperform peers.

Theoretical predictions on this topic are formulated by Thiemann (2016). Here a model of student decision making is developed that explains the different effect of AG by peer effects that have different mechanisms in competitive and non-competitive cultures. Competitive cultures are defined as cultures where social comparison is an important part of the student's utility function. More precisely, competitive students are assumed to compare their own performance with the *best* performance in class, which serves as a reference point in the reference-dependent utility function. In addition, competitive students are assumed to suffer a lot from failures in school, which translates into a high loss aversion. The opposite is true for non-competitive cultures, where social comparison does only weakly influence the students' effort choice and where the *average* performance in class is the reference point of comparison. These assumptions are built on the description of culturally different learning styles by the cross-cultural researcher Hofstede (1986). The hypotheses derived from this model are taken to the data of PISA 2012 in this paper. The aim is to find empirical evidence for the following theoretical predictions. First, we seek general evidence for the existence of an influence of culture on the effect of AG on student performance. Second, predictions on the performance of students in

competitive cultures can be derived from the model. In the simple case with linear utility from Thiemann (2016) comprehensive schooling yields a higher average performance than AG in competitive cultures. In these cultures students have high reference points, such that comprehensive schooling provides all students with the motivating force of a high reference for comparison. In a system with AG, where high-ability students are sorted into a high track and low-ability students into a low track, this positive effect is restricted to the students in the high track. Assuming non-linear utility functions, thereby modeling diminishing sensitivity with respect to the reference point, changes this result. This assumption takes into account the hypothesis that being just below the reference point induces a higher motivation than being further away. In comprehensive schools low-ability students would thus not experience much motivation if they compare with the best student whose performance is too high to be reached. Classes of rather homogenous abilities would then be preferred. This view is also supported by an extension to the model including a participation constraint in Thiemann (2016), which states that students choose not to participate in competition (do not perform at all), if their utility from optimal performance is negative. This extension takes into account that many students would opt out of competition in order to avoid a high loss of utility evoked by loss aversion with respect to a very high reference point. This problem is particularly relevant in comprehensive schools where classes consist of heterogeneous abilities. Competitive students with low ability easily give up in these classes, since the reference point is too far away. Finding evidence for AG being beneficial in competitive cultures would thus support the idea of diminishing sensitivity and participation constraints.

Third, there are predictions for students from non-competitive cultures. The linear model predicts AG to yield a higher average performance than comprehensive schooling in non-competitive cultures. Since students' reference point is the average performance in class, AG can be better at motivating high-ability students since their reference point is higher in a high track than under comprehensive schooling. This effect may on average outweigh the negative effect of AG for low-ability students. The impact of diminishing sensitivity and participation constraints would not change this result, since both assumptions work in general in favor of AG.

Fourthly, theory from Thiemann (2016) predicts that the overall variance of student achievement increases under AG. This is because AG is, both in competitive and non-competitive cultures, detrimental for low-ability students, but beneficial for high-ability students at least in the linear model. If we find evidence for higher variance under comprehensive schooling, this might be evidence for diminishing sensitivity or participation constraints.

Where the theory underlying this paper can describe social preferences and the mechanisms of peer effects in different cultures very precisely, the reality is much more complex. Preferences and likewise culture are not directly observable. Education can be viewed as a black box, where educational inputs (spending, class size, ability grouping) go in and culture-specific outputs are produced. This paper tries to open parts of this black box by using a survey question from the World Values Survey (WVS) (Inglehart, 2014) to derive a measure for country-level competitive preferences.

The theoretical predictions are tested by estimating a typical education production function. This function explains student achievement by multilevel variables: Student background and family information, school characteristics and country specific factors. The empirical estimation of this function uses PISA 2012 math data including roughly 250.000 student observations from 34 countries. The regressor of interest is a measure for AG, which is based on school principals' reports within the PISA study on whether the school groups math classes according to student ability. This school level variable on AG is interacted with the mentioned country level indicator for competitiveness from the WVS. In a least squares approach including country fixed effects the average effect of AG on performance, holding all other factors constant, is estimated. Furthermore, quantile regressions are performed to test the effect of AG across the conditional achievement distribution of students. This also yields insights on the effect of AG on the overall variance of student achievement. As a robustness check an instrumental variable (IV) approach is performed to control for possible endogeneity of the AG variable. This concern exists because of possible student self-selection into schools that perform a certain grouping policy.

The analysis of the PISA 2012 data shows, first and foremost, that culture *does* matter for the effect of AG on student performance. According to the estimation results show students in competitive cultures benefit from AG, whereas students in non-competitive cultures perform lower if they are grouped according to ability. This holds for all students along the conditional achievement distribution, only that students at the tails are generally less affected than those closer to the median. The effect of AG on the variance of achievement is not significantly different from zero in either culture. The IV approach proves to be unnecessary, since endogeneity of the AG variable can be rejected.

The remainder of this paper is organized as follows: Section 2 provides an overview of the related literature. In Section 3 the data used for the analysis is described in detail. In Section 4 the estimation method is outlined. Section 5 reports estimation results. Section 6 provides the IV approach and Section 7 further robustness checks. Section 8 concludes.

2 Related Literature

The question of how AG (sometimes also called *ability streaming* or *ability tracking*) affects students' performance has occupied researchers since the early 20th century. Especially in the USA and the United Kingdom economists have tried to estimate the effect of AG on performance using small student samples from grouped and ungrouped schools. An early literature review is provided by Slavin (1990). The evidence is very mixed, but mostly no strong effect of AG has been found. Since the 1990s bigger data sets are available which has given rise to new approaches in finding an effect of AG. A more recent literature review is provided by Meier and Schütz (2007). There are roughly three strands of literature that empirically analyze the effects of AG: First, there are many studies from the USA that exploit the variation of AG policies within and across American High Schools (e.g. Hoffer, 1992; Argys et al., 1996; Betts and Shkolnik, 2000). Second, there is a strand of literature that uses data from international achievement tests to analyze differences across countries that differ in their national tracking policies (e.g. Ammermüller, 2005; Hanushek and Woessmann, 2006). Third, some studies exist that exploit data from policy reforms and institutional changes in a country's school system (e.g. Pekkarinen et al., 2009; Galindo-Rueda and Vignoles, 2007)). The approach used in this paper combines the first two strands, since effects of AG at the school level are examined, while using international achievement data that includes a variety of countries. To the best of our knowledge there is no empirical literature on the effect of culture on outcomes in education in combination with the effect of AG.

The US studies that analyze AG policies across and within schools struggle with the problem of selection bias. The students' school choice and thus track placement might be affected by unobserved student characteristics such as innate ability, motivation or socio-economic factors. Researchers have developed different strategies to overcome this problem. Hoffer (1992) uses the Longitudinal Study of American Youth (LSAY) to examine the effect of AG on achievement growth from seventh to ninth grade. To overcome criticisms of selection bias Hoffer employs a propensity score approach. He runs a probit regression to predict the probability of high or low track placement for every student and then estimates the effect of actual group placement for different quintiles of these probability distributions. Hoffer does not find a significant effect of grouping on overall average achievement, but finds a moderate positive effect for students in the high group and a stronger negative effect for students in the low group.

Argys, Rees, and Brewer (1996) estimate a selection model to overcome the selection bias problem. They use the US National Education Longitudinal Survey to estimate the

effect of AG on the growth of students' math test scores from 8th to 10th grade. The first-stage of their approach is a multinomial logit model, where track placement for every student is predicted by usage of the following instruments: the racial ethnic make-up of the student body, the region in which the school is located and an indicator for whether the school is located in an urban, suburban or rural community. From these regressions they calculate selectivity correction terms (inverse Mills ratios). In a second stage they include these terms in education production functions that they estimate separately for every track (honors, academic, vocational). The predicted mean achievements are then compared to mean achievement in a heterogeneous class. They find that students in lower tracks would gain from de-tracking, while students in higher tracks would lose. Overall de-tracking would decrease average test scores by 2 %. The Argys et al. (1996) approach is criticized by Figlio and Page (2002), who remark that no evidence on the exogeneity of the instruments is provided.

Betts and Shkolnik (2000) control for unobserved innate ability and motivation by using information on the ability level of the class provided by the teacher. Achievement data is from the LSAY. They do not find an effect of AG on overall achievement, but find that low-ability students are not affected, middle ability students are harmed and high-ability students gain. As a robustness check they estimate a selection model comparable to Argys et al. (1996) using as instruments the percentage of black students in the school, the percentage of students who receive full federal lunch assistance and students' test score relative to the average for his or her grade.

Figlio and Page (2002) use the same data set as Argys et al. (1996) to determine the effect of AG on achievement growth from 8th to 10th grade. They divide the student achievement distribution from 8th grade into top, middle and bottom third and estimate separate regressions for each group. They include a dummy on whether the principal reported that the school applies AG, but find no significant effect in any subgroup. To overcome selection bias, they also estimate a two-stage-least-squares approach using as instruments: the number of schools in the region, the fraction of Reagan voters in the region and the number of academic courses required for state graduation. They only use the interactions of these variables as exogenous instruments to ensure that they are not correlated with achievement. Evidence from this approach suggests that AG has a positive effect on the bottom third and a slight negative effect on students in the top third.

Just like the estimations in this strand of literature, also our estimation might be affected by the problem of selection bias, since school level data on AG is used. In line with Figlio and Page (2002) an instrumental variable approach is employed, contributing to

the literature by suggesting as instrument the number of schools that the given school competes with. This strategy proves to be unnecessary since PISA data provides such a rich set of student background variables that renders the problem of unobserved student characteristics nonexistent.

The second strand of literature uses international achievement studies such as PISA, TIMMS (Trends in Mathematics and Science Study) or PIRLS (Progress in International Reading Literacy Study) to determine the effect of AG. These studies usually define AG on a country level, using different measures such as years spent in tracks, share of students in vocational tracks, the timing of tracking or simply a dummy that indicates whether the country has a grouping policy. Using country-level data comes with the problems of a lack of observations and the difficulty of controlling for all institutional and cultural differences between countries. Ammermüller (2005) tries to overcome this problem by estimating difference-in-difference effects using primary school data from PIRLS and secondary school data from PISA for 12 countries. He can thus cancel out all institutional and cultural country specific effects that do not change over schooling time. His focus is on the question of how changes in institutional variables such as AG influence the strength of the effect of family background variables on achievement. Measuring AG by the number of schools or tracks available to students in secondary schooling, he finds that this variable in combination with parents' education and origin has a positive effect on achievement.

Hanushek and Woessmann (2006) follow a similar approach in also estimating difference-in-difference effects, thus exploiting the fact that all countries that have an AG policy only start sorting after primary schooling. They regress secondary school test scores from TIMMS and PISA on matched primary school test scores from PIRLS and TIMSS. The matching of the tests produces different data sets with 18-26 countries depending on wave and subject. Including a dummy indicating whether the country has a tracking policy they find evidence of a weak negative effect of AG on average performance and a stronger positive effect on inequality, measured by the standard deviation of achievement and differences between percentiles.

Brunello and Checchi (2007) use data from different sources to measure the effect of family background, measured by parental education, in combination with AG on outcome variables for young adults such as earnings, employment, educational attainment and literacy. Their data set spans over several years and includes 12-25 countries depending on the outcome variable. To control for country specific effects they include country by cohort dummies. They find that the effect of family background becomes stronger with AG. Another result is that AG causes a stronger dispersion of earnings.

This paper can contribute to this literature by using school-level data, that has the advantage that it has a much higher variance in the AG variable and country fixed effects can be included to control for unobserved country specific factors.
The third strand of literature investigates policy reforms to learn from institutional changes. There are two papers by Galindo-Rueda and Vignoles (2007) and Manning and Pischke (2006) that investigate the gradual change from a selective to a comprehensive school system in England and Wales in the 60s and 70s. Whereas Galindo-Rueda and Vignoles find that a selective school system favors high-ability students, Manning and Pischke (2006) do not find any significant effects. Pekkarinen et al. (2009) take a look at the Finnish reform from a two-track system to a comprehensive school system that took place gradually in 1972-1977. The major finding from their difference-in-difference approach is that the reform reduced inequality, as proven by a significant drop in the intergenerational income elasticity by 23%.
Overall, positive effects of AG are usually assigned to the channel of better targeted pedagogy (see Cortes and Goodman, 2014; Duflo et al., 2011), while the channel of peer effects is made responsible for the positive effects on high skilled students and negative effects on low skilled students (see Argys et al., 1996; Hoffer, 1992). The mixed evidence from past research has therefore several reasons. First, there is a lack of disentanglement of the channels through which the effects of AG work. Second, there are many different empirical approaches and different definitions of AG. Third, many of the reviewed studies are based on different subject pools from different countries and therefore different cultures. This paper contributes to the first point by aiming at explaining effects through the channel of peer effects only. Most importantly we also contribute to the third point by showing that effects of AG differ between cultures of competitiveness and thereby provide an explanation for the mixed evidence from past research.

3 Data

3.1 Student Achievement Data

Student achievement is measured using data from the 2012 PISA study. In the 2012 wave the acquired knowledge of about 510,000 15-year-old students from 65 countries is assessed in three key areas: reading, mathematics, science and problem solving. In the focus area, mathematics, students solved paper and pencil test questions that assess their capacity to formulate, employ and interpret mathematics in a variety of contexts. The test lasts about 2 hours and subsequently students have to fill in a background questionnaire. The

individual student assessment is measured on a scale that is based on a mean for OECD countries of 500 points and a standard deviation of 100 points that were set in PISA 2003 when the first PISA scale was developed (OECD, 2013b). The sampling procedure of PISA is a two-stage sampling design. For each country first a sample of schools is selected from a complete list of schools containing the student population of interest. Then, a simple random sample of 35 students from the 15-year-old student population is drawn from within the selected schools (OECD, 2014). The principal of the selected school is asked to complete a questionnaire on school characteristics, generating a data set including student and school level variables.

The data used for this paper includes only 34 of the 65 PISA countries. The number of countries is reduced for two reasons. First, cultural data is not available for all countries.[1] Second, we only include countries that have a comprehensive school system on a national level. The variable used as an identifier for AG in this paper is a school level variable that yields information on whether classes within the school are grouped according to ability (see Section 3.3 for more details). This approach yields more variance and observations than comparing tracked and comprehensive school systems on a country level. Since PISA does not take into account that schools might be part of a nationally tracked school system, the variable on AG within schools might be biased. The school might already be a selection of low or high-ability students, if the whole school is part of a nationally tracked system. The effect of additional grouping on performance in these schools is not the same as in a comprehensive school system. To solve this problem countries with a tracked school system are excluded from the estimation.[2] Furthermore, we delete all observations of first or second generation immigrant students. Since we assume that competitiveness is a value that is transmitted from parents to their children, we cannot assign national culture to immigrants. This results in a loss of 27,157 observations. Table 1 lists the 34 countries, the number of schools and students included in the analysis. Even though the PISA sample is generally biased towards developed countries, the latest test from 2012 includes a wide variety of cultures, including non-OECD countries from South-America and Asia. Altogether 10,588 schools and 251,972 student observations are included.

The reasons for using PISA 2012 data are, first, that it provides a huge database covering a large number of countries to ensure that there is enough variation in terms of culture. Second, the PISA 2012 questionnaire contains a question on AG that fits the purposes

[1] These countries are: Costa Rica, Cyprus, Denmark, Greece, Ireland, Iceland, Israel, Liechtenstein, Macao-China, Montenegro, Portugal, Shanghai-China, Tunisia, United Arab Emirates.

[2] These countries are: Germany, The Netherlands, Belgium, Turkey, Austria, Switzerland, Luxembourg, Bulgaria, Romania, Hungary, The Czech Republic, Uruguay, Singapore, Korea, Italy, Croatia. Information on tracked school systems is taken from the OECD (2013a, p.78).

Table 1: Student and School Observations by Country

code	country	schools	students	code	country	schools	students
ALB	Albania	187	4,246	LTU	Lithuania	209	4,470
ARG	Argentina	220	5,437	LVA	Latvia	203	3,964
AUS	Australia	759	11,738	MEX	Mexico	1,453	33,050
BRA	Brazil	648	16,861	MYS	Malaysia	163	5,076
CAN	Canada	862	17,435	NOR	Norway	186	4,038
CHL	Chile	218	6,737	NZL	New Zealand	155	2,842
COL	Colombia	323	8,565	PER	Peru	238	5,993
ESP	Spain	862	22,338	POL	Poland	166	4,194
EST	Estonia	202	4,359	QAT	Qatar	143	4,718
FIN	Finland	301	7,433	RUS	Russia	224	4,614
FRA	France	200	3,528	SRB	Serbia	138	3,950
GBR	Great Britain	473	10,903	SWE	Sweden	207	4,033
HKG	Hong Kong	147	3,054	TAP	Taiwan	163	6,016
IDN	Indonesia	206	5,533	THA	Thailand	239	6,571
JOR	Jordan	225	5,960	TUN	Tunisia	150	4,303
JPN	Japan	190	6,293	USA	USA	155	3,881
KAZ	Kazakhstan	211	4,885	VNM	Vietnam	162	4,954
					Total	10,588	251,972

of this study. Since the focus of the PISA 2012 study was on mathematics, the question on AG asks specifically for grouping in *mathematics* classes. This is ideal, since in the analysis only the achievement data from the mathematics test is used. This is because math data is generally viewed as being most comparable across countries (Hanushek et al., 2013). Third, PISA data has the huge advantage that test scores are comparable across all students, schools and countries. This is important, since comparing school grades of grouped and ungrouped students might otherwise be biased because of different grading practices. Figure 1 shows 2012 mean performance in mathematics for the 34 countries in the sample. Mean performance is highest in East-Asian countries and lowest in South-American and South-East-Asian countries.

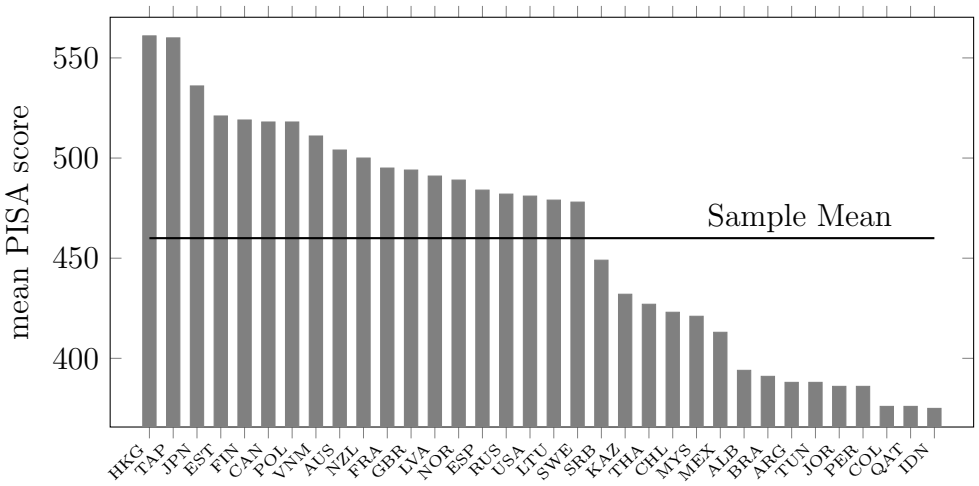

Figure 1: Mean Score in Mathematics by Country in PISA 2012 (OECD, 2013b)

3.2 Measure of Culture of Competitiveness

Culture is a highly subjective matter and hence hard to measure in numbers. In recent years international surveys have tried to make culture comparable across countries. Data is thus only available on a country level, which is generally justified by the fact that people from the same country share important determinants of the development of culture such as language and history. In this study a measure for a country's competitiveness is derived from a question from the WVS. The WVS is a global network of social scientists studying changing values and their impact on social and political life. The survey started in 1981 and consists of national surveys conducted in almost 100 countries using a common questionnaire. Random sampling is used in the countries to obtain representative national samples (Inglehart, 2014). From the WVS answers to the following statement are taken: "*Competition is good. It stimulates people to work hard and develop new ideas" vs. "Competition is harmful. It brings the worst in people*". Participants were asked to place their view about this statement on a scale from 1 to 10, where 1 means "*competition is good*" and 10 means "*competition is harmful*". 107,466 people were interviewed between 1989 and 2012. The data from all waves is aggregated on a country level and a simple average per country is calculated. The $Comp_c$ index is created by normalizing the data to take on numbers between 0 and 10, and reverse coded so that 10 is the most competitive country and 0 the least competitive. It is assumed that school children's competitiveness is captured by this aggregated measure since cultural values are shared by large groups or nations and are transmitted from parents to their children through generations. Looking at breakdowns of the data by age shows that the measure hardly changes if we only take the average of young participants or from old people. To confirm the time persistence of this cultural values, we calculate a $Comp_{ct}$ index per wave and run a panel data regression of the $Comp_{ct}$ index on country and time dummies. Performing F-tests on the time dummies proves that these are insignificant (p-value: 0.13). According to the created index $Comp_c$, competitive countries are those from Eastern Europe, the Balkan countries, the USA and Australia. Non-competitive countries are those from South-America and Western Europe. Asian countries are moderately competitive. For a full ranking of countries see Appendix A.

To investigate further what factors determine the $Comp_c$ index, we calculate relevant correlations with country-level variables and discuss existing literature using the same measure. First, note that the correlation with the mean country score of PISA 2012 is negative and non-significant (-0.215), indicating that a competitive culture is not associated with a higher average achievement at school (see Figure 2). Hayward and Kemmelmeier (2007) examine the structural and cultural roots of competitive attitudes

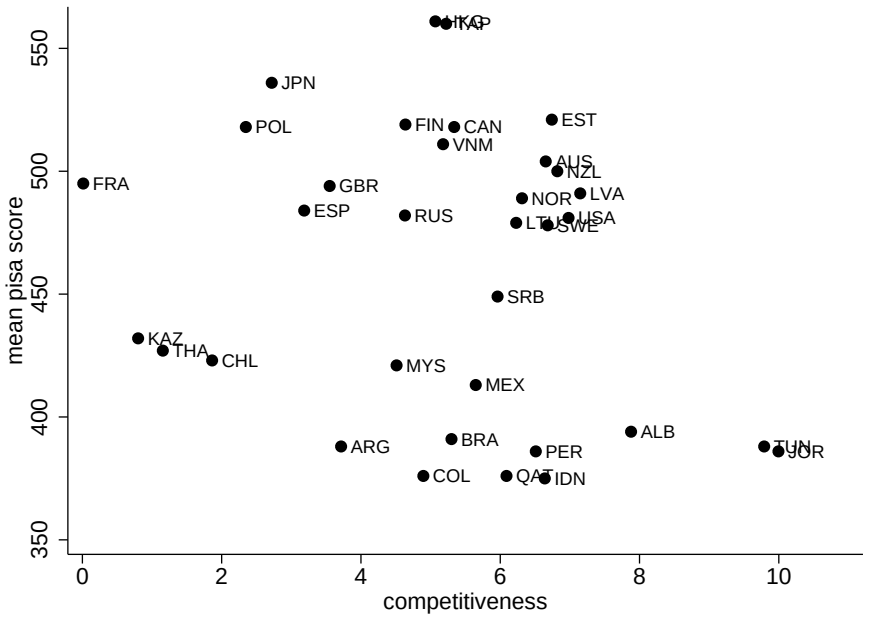

Figure 2: Relationship of Mean PISA Test Score and Competitiveness according to the WVS, Normalized to Values from 0 (Non-Competitive) to 10 (Competitive)

using the $Comp_c$ measure including 81 countries. They find that the index is, if at all, negatively correlated to economic prosperity as measured by the per capita GDP or to economic freedom of a country as measured by the Heritage Foundation's Index of Economic freedom. The only significant finding of Hayward and Kemmelmeier (2007) is that competitive values are consistently correlated to Protestantism across societies as measured by the proportion of Protestants of the national population. According to the authors the Protestant culture is a value system that promotes the principles of free-market enterprise, and is hence likely to promote a competitive mindset. In our sample these are the Anglo-Saxon countries as well as Scandinavian countries. Hayward and Kemmelmeier (2007) find no correlation to individualism as opposed to collectivism measured by Hofstede (1984), who undertook an extensive survey about values at the workplace. In addition we calculate the correlation with the "Masculinity vs. Femininity" (MAS) measure developed by Hofstede (1984), which measures performance orientation as associated with masculine societies vs. cooperation orientation as associated with feminine societies. The correlation with this index is also small and insignificant (0.072). We argue that $Comp_c$ does thus not capture values measured by MAS such as performance orientation or free-market orientation and prosperity as measured by the GDP. Instead the WVS question used for $Comp_c$ does specifically mention the word "competition", so that it captures the aspect of social comparison.[3]

[3]Conducting the same analysis as done in this paper with MAS instead of COMP, yields insignificant

Guiso, Sapienza, and Zingales (2003) study the impact of religion on economic attitudes, also using the $Comp_c$ measure under consideration in this paper. They find that Catholics and Protestants are in favor of competition, whereas Muslims and Hindus are strongly against it. Hindu countries (THA, VNM) and some Muslim countries (MYS, IDN) also score low in our sample. However, Jordan and Tunisia as non-Asian, but Muslim countries are obvious outliers, as well as France being Catholic but non-competitive.

3.3 Measure of Ability Grouping

The policy of AG implies that students are sorted into groups based on their ability or past achievement. These groups are then taught on different levels of difficulty. There are, however, several forms of AG that differ in their rigidity: First, there is the most rigid form: *countrywide ability tracking*. This means students are separated into different schools, usually based on achievement in primary school. Secondary schooling is then organized in two or three different tracks (schools). In this form of AG students are completely sealed off from students with different abilities. Second, there is *between-class grouping*, where students are separated into different classes within a school based on ability levels. And third, there is *within-class grouping*, where a class is divided into groups based on ability and achievement. This is commonly accomplished by assigning every member of the class to a particular group that they will be taught with during instruction in a particular subject. This is the least rigid form, since students still know and observe students with heterogeneous abilities within their class.

The strongest effect of AG is expected when students are grouped into different schools, so that reference point formation is only possible within the schools. Since this form of AG is implemented on a country level there is only little scope for regression analysis. Too few observations are available and a lot of other country-specific factors are likely to confound the analysis. Effects of AG are hardly ever found with this approach (Hanushek and Woessmann, 2006). Still, the same mechanisms should be at work when considering between-class grouping, the second most rigid form. If significant effects are found here, the effects in countries with rigid track formation should be even stronger. Considering between-class-grouping enables us to conduct the analysis on a school level, yielding many more observations and variation.

results, indicating that the effect of AG does not depend on values measured by MAS. There is, however, a positive correlation of MAS with the average country score from PISA 2012, which illustrates the performance orientation measured by MAS.

"*Schools sometimes organize instruction differently for students with different abilities and interests in mathematics.* **Which of the following options describe what your school does for 15-year-old students?**" Please tick one box per row.

	For all classes	For some classes	Not for any classes
a) Mathematics classes study similar content, but at different levels of difficulty.	☐	☐	☐
b) Different classes study different content or sets of mathematics topics that have different levels of difficulty.	☐	☐	☐

Figure 3: Question on Ability Grouping in the PISA 2012 School Questionnaire

The PISA school principal questionnaire includes a question on ability grouping that is shown in Figure 3. From this question the variable AG_{sc} is constructed. This variable has the following six categories: (0) "not for any classes" for both a) and b); (1) "for some classes" for either a) or b) and "not for any classes" for the other; (2) "for some classes" for both a) or b); (3) "for all classes" for either a) or b) and "not for any classes" for the other; (4) "for all classes" for either a) or b) and "for some classes" for the other; (5) "for all classes" for both a) and b). Of all schools in the sample 16% have no AG (category 0), 13% are in category 1, 29% have some AG as in category 2, 16% of schools are categorized into 3, 14% in 4 and 11% of schools group all classes as in 5. Figure 4 shows the percentage of schools in the respective category of the variable AG_{sc} for all 34 countries included in the sample. The variable shows sufficient variation within and between countries. Remeber that only countries with a countrywide comprehensive school system are included in the sample as explained in Section 3.1. Countries with a relatively high percentage of grouped classes are (traditionally) English-speaking countries like Great Britain, the USA, New Zealand and Australia. Relatively little AG can, for example, be found in Scandinavian countries. The correlation between the amount of AG in a country (mean of AG_{sc} by country) and the culture of competitiveness is positive, but rather low and not significant (0.24).

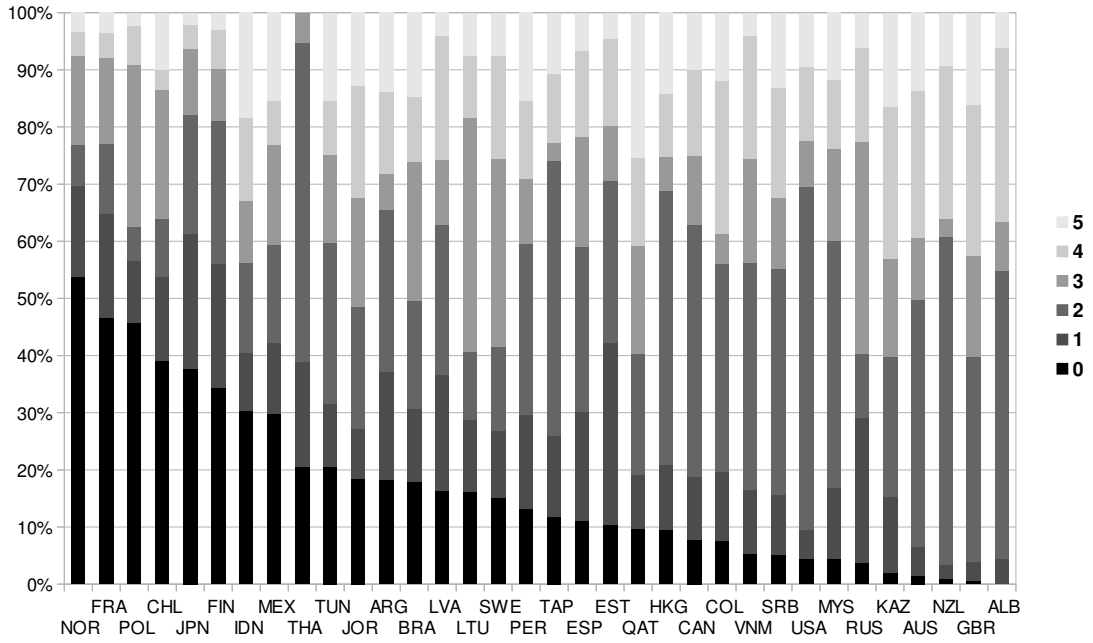

Figure 4: Share of Schools in a Country according to Categories of AG_{sc}

3.4 Control Variables

A standard set of control variables at the student and school level as found in many recent publications using PISA data is included (see e.g. Hanushek, Link, and Wößmann, 2013). In addition, some context related variables that might be correlated with our variable for AG are also added. Table 2 describes all control variables used in all following estimations.

Table 2: Description of Control Variables

Variable Name	**Description**
Student level:	
Age	Age of the student in years
Female	Dummy=1 if student is female
Grade Repetition	Dummy=1 if student ever repeated a grade
Grade	Grade of the student compared to modal grade for 15-year-olds in the country
Other Language at Home	Dummy=1 if student speaks a different language than the test language at home
Parents' Education	Highest completed level of education of both parents with categories: None (1), Primary School (2), Lower Secondary (3), Upper Secondary 1 (4), Upper Secondary 2 (5), University (6)
Books	Books at the home of the student (excluding school textbooks) with categories: 0-10 (1), 11-25 (2), 26-100 (3), 101-100 (4), 201-500 (5), more than 500 (6)
Index of Socio-Economic Status (HISEI)	Index of the parents' socio-economic status, ranging from 0-100, taking into account their occupation and wealth

Table 2: (continued)

Variable Name	**Description**
Class Size	Class size of the student's test language class
School level:	
Number of Students	Total student enrollment at the school
Private School	Dummy=1 if the school is a private school
Government Funding	Share of funding by the government
School Location	School location with categories: Village (1), Small Town (2), Town (3), City (4), Large city (5)
Math-Teacher Shortage	Dummy=1 if principal reports a shortage of math teachers
Student-Teacher-Ratio	Ratio of number of students to number of math-teachers at school
School Autonomy	Index on how much autonomy the school has regarding school budget, hiring and firing of teachers, teacher salary, courses offered etc.
Admission by Ability	Indicator on whether the school admits students based on academic record with categories: Never (1), Sometimes (2), Always (3)
Same Textbook	Dummy=1 if the school uses the same mathematics textbook for all classes

It is controlled for many factors that might determine whether a school practices AG or not, for instance the total number of student enrollment, school location, the type of school (private vs. public) and the share of government funding. Also racial and socioeconomic heterogeneity of a schools student body influence a schools decision to group (VanderHart, 2006). Including variables that control for this (e.g. *Other Language at Home*, *Books*, *Parents' Education*) ensures that there is no omitted variable bias, in the sense that the indicator for AG just picks up the effect of one of these variables. Controlling for whether the school admits students based on their prior achievement (*Admission by Ability*) is also important, since the student body at a school that undertakes this policy is a selection of high-ability students and AG in such a school probably has a lower effect.

The effect of AG on performance is not only driven by peer effects, but also by other factors like more appropriately paced instruction, smaller class size and focused curricula in ability segregated groups. For some of these factors it is controlled for by variables within the school characteristics vector, e.g. *Class Size*, which is usually smaller in schools where AG is used. Also, we control for *Same Textbook*, which indicates whether the school uses better suited curricula for different ability groups. Furthermore, we argue that if significant effects for an interaction of AG with competitiveness are found, there must be peer effects at work, since the variable $Comp_c$ is defined by social comparison. Table 3 shows correlations of AG_{sc} with important control variables. All the correlations are highly significant, but low in size.

Table 3: Pairwise Correlations of Ability Grouping and Selected Control Variables

	Ability Grouping		Ability Grouping
Class Size	-0.0536***	Number of Students	0.0173***
Books	-0.0370***	Private School	0.0139***
Index of Socio-Economic Status	-0.0065***	Admission by Ability	-0.0409***
Government Funding	0.0374***	Same Textbook	-0.0826***
School Location	-0.0135***		

Notes: Weighted by students sampling probability. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

More than fifty percent (58%) of the students have one missing value in at least one of the reported control variables. There is no pattern of missing values, but the values seem to be missing at random (MAR) in a non-monotone manner. Dropping all students with missing values would result in a substantial loss of observations and would lead to biased coefficients. As a solution missing data is imputed using the data of students with non-missing data as proposed by Woessmann (2003) and Ammermüller (2005). See Appendix B for a detailed description of the imputation technique. Appendix C provides summary statistics (mean and standard deviation) of student achievement and all imputed control variables.

4 Estimation Technique

The underlying model is an education production function framework, which typically explains student achievement by variables on the individual, the school and the country level (see e.g. Woessmann, 2003) resulting in a multi-level model. This model is augmented by the measure of competitiveness.

$$A_{isc} = \alpha + \beta_1\ AG_{sc} + \beta_2\ Comp_c + \beta_3\ AG_{sc} \times Comp_c + \mathbf{FB}_{isc}\gamma + \mathbf{S}_{sc}\delta + \mathbf{C}_c\kappa + \epsilon_{isc} \quad (1)$$

The dependent variable A_{isc} is math achievement of student i in school s and in country c as measured by PISA 2012.[4] The variable AG_{sc} is the indicator for AG as described in Section 3.3. The variable $Comp_c$ is the country level indicator for competitiveness as

[4]PISA does not offer a single variable for student achievement, but 5 plausible values. Plausible values are random values drawn from a mathematically computed distribution of students' ability based on their test results and provide better estimates at the population level. Instead of one, there are thus five regressions to be computed for five different dependent variables. Results for coefficients and standard errors are averages of the results from the five plausible value regressions.

described in Section 3.2. To test whether the impact of AG varies with the competitiveness of students an interaction of AG_{sc} and $Comp_c$ is included. The vector $\mathbf{FB}_{isc}$ is a vector of the family background variables, $\mathbf{S}_{sc}$ a vector of the school characteristics and $\mathbf{C}_c$ is a vector of country characteristics. The error term is composed of errors at the individual student level, at the school level and at the country level:

$$\epsilon_{isc} = \eta_c + \eta_{sc} + \eta_{isc} \tag{2}$$

Here the country-specific error term η_c includes a set of cultural and educational factors for country c that cannot be measured, η_{sc} is a school-specific and η_{isc} an individual-specific error term. Since the purpose here is to find effects at the school level, country fixed effects μ_c can easily be included to control for unobserved country-specific factors, i.e. get rid of η_c. This also eliminates the variable $Comp_c$ because of perfect multicollinearity with the fixed effects. However, $Comp_c$ can stay in the interaction, which varies on a school level.

$$A_{isc} = \alpha + \beta_1 AG_{sc} + \beta_3 \; AG_{sc} \times Comp_c + \mathbf{FB}_{isc}\gamma + \mathbf{S}_{sc}\delta + \mu_c + \epsilon_{isc} \tag{3}$$

The error term is now only composed of errors at the individual and at the school level, η_{sc} and η_{isc}. It is not possible to include school fixed effects, since this would eliminate the AG_{sc} variable. However, a wide set of school level variables is included, assuming that there are no unobserved school-specific effects left that are correlated with AG_{sc}. Despite the country fixed effects we still need the assumption of no unobserved cross-country heterogeneity that is related to the effect of AG on achievement for the identification of Equation (3). The only two channels discussed in the literature that determine how achievement is influenced by AG are peer effects and more appropriately paced instruction (Hanushek and Woessmann, 2006). Since the latter channel is unlikely to vary with culture, we assume that the coefficient β_3 captures the influence of culturally varying peer effects.

Equation (3) is estimated using ordinary least squares (OLS). To take into account the clustered nature of the data, where students are nested within schools and the schools are nested within countries, cluster-robust standard errors are used at the highest, namely the country level. The sampling design of PISA is not completely random, which is why weights are used for every student consisting of the school weight and within-school weight to account for different sampling probabilities. The complex survey design of PISA also makes it necessary to use replication methods for computing the sample variance. PISA suggests Balanced Repeated Replication (BRR) with Fay's modification (OECD, 2005,

pp.23), which is used here accordingly.[5]

The main interest is in the coefficients β_1 and β_3 of Equation (3). The coefficient β_1 can be interpreted as the change in average math achievement, if the variable AG_{sc} increases by 1 category for students in non-competitive countries (i.e. $Comp_c = 0$). The coefficient β_3 is the change in average achievement, if $Comp_c$ increases by one for students that are subject to AG as in category 1.

The regression might still suffer from selection bias, since good students could be attracted by schools that have a system with AG. This problem can be interpreted as an omitted variable bias, with innate ability being the omitted variable. This would result in η_{isc} being correlated with AG_{sc}. If this is the case, we would expect β_1 and β_3 to be positively biased. Some researchers (e.g. Ammermüller, 2005) argue that the problem of omitted ability does not matter in education production frameworks, since many proxy variables for ability are already included (e.g. parents' education, number of books at home, parents' occupation). The omitted variable issue shall still be considered in Section 6 as a robustness check.

5 Results

5.1 Pooled OLS wit Country Fixed Effects

The results from an OLS estimation of Equation (3), with and without the interaction of AG_{sc} with the $Comp_c$ index, are presented in Table 4. The estimated coefficients of all included control variables are given in Appendix C. These coefficients are in line with previous research using PISA (e.g. Woessmann, 2003; Hanushek and Woessmann, 2014). Specification (1) and (3) in Table 4 include the variable AG_{sc} as the ordered categorical variable described in Section 3.3, and specification (2) and (4) in dummy coding with "no classes grouped" ($AG_{sc} = 0$) being the reference category. From the specifications with AG_{sc} in dummy coding we can conclude that these estimations are more meaningful, since the distances between the coefficients on the different categories of grouping are very different.

[5]For regression computing STATA is used with the PV module designed by Macdonald (2014).

Table 4: The Effect of Ability Grouping on Achievement (Pooled OLS)

Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.220	-3.457**		
	(0.596)	(1.441)		
Ability Grouping × Comp		0.615**		
		(0.292)		
AG = 1 (Some Classes Grouped)			-1.660	-10.500*
			(3.355)	(6.109)
AG = 2			-0.942	-21.558***
			(2.620)	(5.753)
AG = 3			0.188	-8.431
			(2.698)	(5.954)
AG = 4			-1.528	-16.579**
			(3.265)	(7.733)
AG = 5 (All Classes Grouped)			-1.878	-18.751**
			(3.709)	(8.689)
AG = 1 × Comp				1.911
				(1.468)
AG = 2 × Comp				4.226***
				(1.206)
AG = 3 × Comp				1.894
				(1.246)
AG = 4 × Comp				3.103*
				(1.662)
AG = 5 × Comp				3.385**
				(1.689)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,558	10,558	10,558	10,558
Student obs.	249,968	249,968	249,968	249,968
avrg. R^2	0.49	0.49	0.49	0.49

Notes: Dependent variable: PISA 2012 math test score. Reference Category is 'no grouping in any classes'. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: age, female, parents' education, hisei, grade, grade repetition, books at home, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, math-teacher shortage, same textbook, admission by ability, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

To interpret coefficients note that the PISA math score was normalized to have a mean of 500 and a standard deviation of 100 across OECD countries in 2003. The first and third specification in Table 4 show no significant effect of AG on achievement when the interaction term with culture is not included. This corresponds to previous research that did not find effects of AG on average performance. However, the specifications that include the WVS measure for competitiveness show that culture *does* matter. Here AG has a significant negative effect in countries with low competitiveness, but a positive effect in competitive countries. For example, from specification (4) we see that in a country scoring 0 on the $Comp_c$ index, AG in all classes (as in category 5) compared to AG in no classes reduces achievement on average by 19 score points. In a country scoring 10

on the $Comp_c$ index AG in all math classes increases average achievement by about 15 score points. Finally, for a medium competitive country of $Comp_c = 5$ there is no effect of grouping in all classes. Specification (4) also shows that already grouping in "some classes" as in category 2 of the variable AG_{sc} has a strong effect. It leads to a decrease of 22 points of average student achievement in non-competitive countries ($Comp_c = 0$) and an increase of 21 points in competitive countries ($Comp_c = 10$). Schools reporting that *some* classes are grouped might, for instance, be comprehensive schools that have remedial classes for particularly bad students or extra math classes for particularly good students. The estimated coefficients are relatively large compared to estimated effects of school inputs in the PISA literature. For example Fuchs and Wößmann (2008) find that 1000 hours of extra instruction time per year lead to an increase in average achievement by 5 score points and that students at a publicly managed school perform on average 19 score points lower than students at privately managed schools.

5.2 Quantile Regression

To test whether low or high-ability students suffer or gain more from AG, quantile regressions with country fixed effects according to Koenker (2004) are run. This enables us to see whether there is heterogeneity in the effects of grouping across the conditional achievement distribution. Since it is controlled for all kinds of family and student characteristics, the conditional achievement distribution should be strongly correlated with innate ability, or more precisely the part of ability that is not correlated to the measured student characteristics (for a similar approach see Woessmann, 2008). We will thus from now on refer to the conditional achievement distribution as the ability distribution. Since it can be assumed that the distribution of innate ability is constant across countries, we do not have to worry about the different achievement distributions in different countries. Quantile regressions estimate the effect of grouping on student achievement for students at different points on the ability distribution. Table 5 reports the coefficients on AG_{sc} and $AG_{sc} \times Comp_c$ in dummy coding for the quantiles ranging from 0.1 to 0.9. Parente and Santos Silva (2013) show that the quantile regression estimators are also consistent when the error terms are correlated within clusters.

For students at all quantiles we find significant negative effects in non-competitive countries and significant positive effects in competitive cultures. Focusing on the coefficients on the $AG = 5$ dummy, which indicates the change in average achievement if all classes in the school are grouped compared to no grouped classes, we see that the effect of AG is biggest for students at the median and becomes smaller the further away we go in each direction. In very non-competitive cultures ($Comp_c = 0$) AG in all classes decreases

Table 5: Quantile Regressions

Variables	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AG=1	-4.474**	-6.798***	-9.565***	-11.143***	-12.873***	-9.226***	-11.423***	-6.747***	-5.529**
	(2.199)	(1.800)	(1.705)	(1.685)	(1.692)	(1.716)	(1.813)	(1.833)	(2.288)
AG=2	-11.793***	-16.344***	-16.695***	-19.346***	-21.184***	-21.969***	-23.391***	-24.219***	-24.319***
	(1.974)	(1.616)	(1.530)	(1.512)	(1.518)	(1.540)	(1.627)	(1.645)	(2.054)
AG=3	-4.880**	-6.572***	-6.752***	-7.103***	-10.642***	-8.205***	-8.212***	-6.324***	-4.553*
	(2.266)	(1.855)	(1.757)	(1.736)	(1.743)	(1.768)	(1.868)	(1.889)	(2.358)
AG=4	-7.331**	-11.384***	-17.530***	-18.095***	-22.570***	-19.798***	-19.887***	-17.919***	-16.118***
	(2.858)	(2.339)	(2.215)	(2.190)	(2.198)	(2.229)	(2.356)	(2.382)	(2.973)
AG=5	-13.210***	-15.748***	-16.925***	-18.020***	-21.482***	-14.768***	-13.483***	-10.524***	-9.262***
	(3.199)	(2.619)	(2.480)	(2.451)	(2.461)	(2.496)	(2.638)	(2.667)	(3.329)
AG=1 × Comp	0.211	1.099***	1.756***	1.984***	2.519***	1.664***	1.912***	0.695*	0.873*
	(0.439)	(0.360)	(0.341)	(0.337)	(0.338)	(0.343)	(0.362)	(0.366)	(0.457)
AG=2 × Comp	2.097***	3.347***	3.595***	3.975***	4.347***	4.392***	4.479***	4.435***	4.760***
	(0.372)	(0.304)	(0.288)	(0.285)	(0.286)	(0.290)	(0.306)	(0.310)	(0.387)
AG=3 × Comp	1.377***	1.823***	1.812***	1.705***	2.360***	1.854***	1.757***	1.097***	0.689
	(0.434)	(0.356)	(0.337)	(0.333)	(0.334)	(0.339)	(0.358)	(0.362)	(0.452)
AG=4 × Comp	1.304**	2.231***	3.335***	3.300***	4.345***	3.941***	3.741***	3.091***	3.219***
	(0.515)	(0.421)	(0.399)	(0.394)	(0.396)	(0.402)	(0.424)	(0.429)	(0.536)
AG=5 × Comp	1.922***	2.665***	3.180***	3.357***	3.841***	2.680***	2.437***	1.706***	1.982***
	(0.563)	(0.461)	(0.436)	(0.431)	(0.433)	(0.439)	(0.464)	(0.469)	(0.586)
Student controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34	34	34	34	34	34
School obs.	10,558	10,558	10,558	10,558	10,558	10,558	10,558	10,558	10,558
Student obs.	249,968	249,968	249,968	249,968	249,968	249,968	249,968	249,968	249,968
Avrg. pseudo R^2	0.22	0.25	0.27	0.28	0.30	0.31	0.32	0.32	0.33

Notes: Dependent variable: PISA 2012 math test score. Quantile regression weighted by students sampling probability. Standard errors given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, books at home, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

achievement by 13 score-points for students with very low ability at the 0.1 quantile and decreases achievement by 9 score-points for high-ability students at the 0.9 quantile. In very competitive cultures ($Comp_c = 10$) AG in all classes increases achievement by 19 score-points for low-ability students at the 0.1 quantile and by roughly the same amount for high-ability students at the 0.9 quantile. The median regression can be viewed as a test of the OLS regression that is robust against outliers (Woessmann, 2008). Here it clearly supports the results of the least-squares regression with coefficients on $AG = 5$ and $AG = 5 \times Comp$ (all classes grouped) being a bit bigger.

Figure 5 shows the impact of $AG = 5$ (all classes grouped compared to no classes grouped) on achievement in score-points for non-competitive countries ($Comp_c = 0$, in light grey) and competitive countries ($Comp_c = 10$, in dark grey) across the quantiles. It shows both the estimates including all the control variables from previous estimations (see Table 5) as well as from quantile regressions without the control variables that might be proxies for

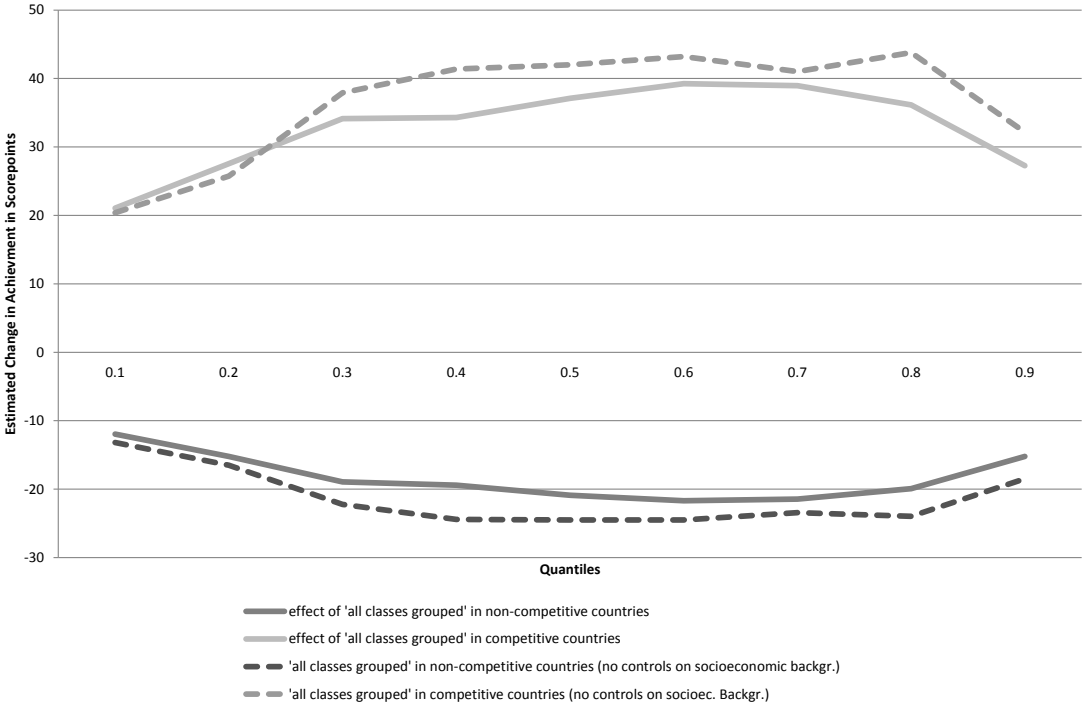

Figure 5: Estimated Effect of "All Classes Grouped" at Different Quantiles of the Conditional Achievement Distribution for Competitive and Non-Competitive Cultures

ability[6] (dashed lines). This is done as a robustness check to ensure that the correlation of the conditional achievement distribution with innate ability is not eliminated by these control variables. The conditional achievement distribution from estimations excluding these control variables should then be highly correlated to innate and nurtured ability.
Figure 5 shows that the influence of AG is generally smaller at the tails of the ability distribution. A possible explanation for this result is that medium-ability students are confronted with the biggest change in their social position when they are sorted into ability based groups. While they are mediocre under comprehensive schooling they are either among the best or worst students in a two-track system. Without controlling for variables on socioeconomic background the effects are generally the same, but slightly bigger. This might be because also nurtured ability is important for the effect of AG on achievement. Another explanation is that these socioeconomic variables are correlated with AG_{sc}, so that the coefficient on AG_{sc} in the regressions without these controls catches some of their effects.
The results from the quantile regressions also yield insights on the effect of grouping on the variance or inequality of achievement. Since only little variation is found across

[6] i.e. without *books at home*, *hisei* and *parents' education*

the quantiles, variance effects of AG should be small. In non-competitive countries the variance in grouped schools is larger than in comprehensive schools, since low-ability students lose more from grouping than high-ability students. In competitive countries the variance is also slightly bigger under grouping, since high-ability students gain more than low-ability students. Also a regression using the standard deviation of achievement per school as dependent variable supports the result that between-class grouping has little influence on the inequality of student achievement (see Appendix E).

6 Instrumental Variables

There is a possibility that the variable AG_{sc} is endogenous. This is because school choice of students (or their parents) might be affected by whether a school does or does not group by ability. For example, good students might be attracted by schools that have groups for high-ability students, since this gives them the opportunity to study at a higher difficulty without being slowed down by low-ability students. In this case the above estimates for AG_{sc} would be biased upwards. This problem can also be interpreted as an omitted variable bias, as in Betts and Shkolnik (2000), with innate ability being the omitted variable. Since innate ability is probably positively correlated with AG_{sc}, an endogeneity problem arises. In order to address this problem an instrumental variable approach is suggested using as an instrument a variable that yields information on whether students have a choice between different schools.

The instrument suggested is data from a question from the PISA 2012 school questionnaire about how many schools the school is competing with in the region. From this question a variable $Schoolcomp_{sc}$ is constructed that takes on the value 0 if the school is not competing with any other school, 1 if the school is competing with one other school and 2 if there are two or more schools the school competes with. Figure 6 illustrates that there is a lot of variation in this variable between and within countries. Naturally more availability of schooling is found in countries that are more densely populated.

A positive correlation between $Schoolcomp_{sc}$ and AG_{sc} is expected for two reasons. First, the availability of schooling in the region is a natural predictor of self-selection since no selection can take place when students do not have a choice between schools. Therefore the effect of AG on students that do not have a choice between schools can be compared with those that have a choice. Second, school competition might also affect a schools decision to group classes or not to group. If a school is competing for students with other schools, it might rather offer ability-grouped classes in order to attract high-ability students.

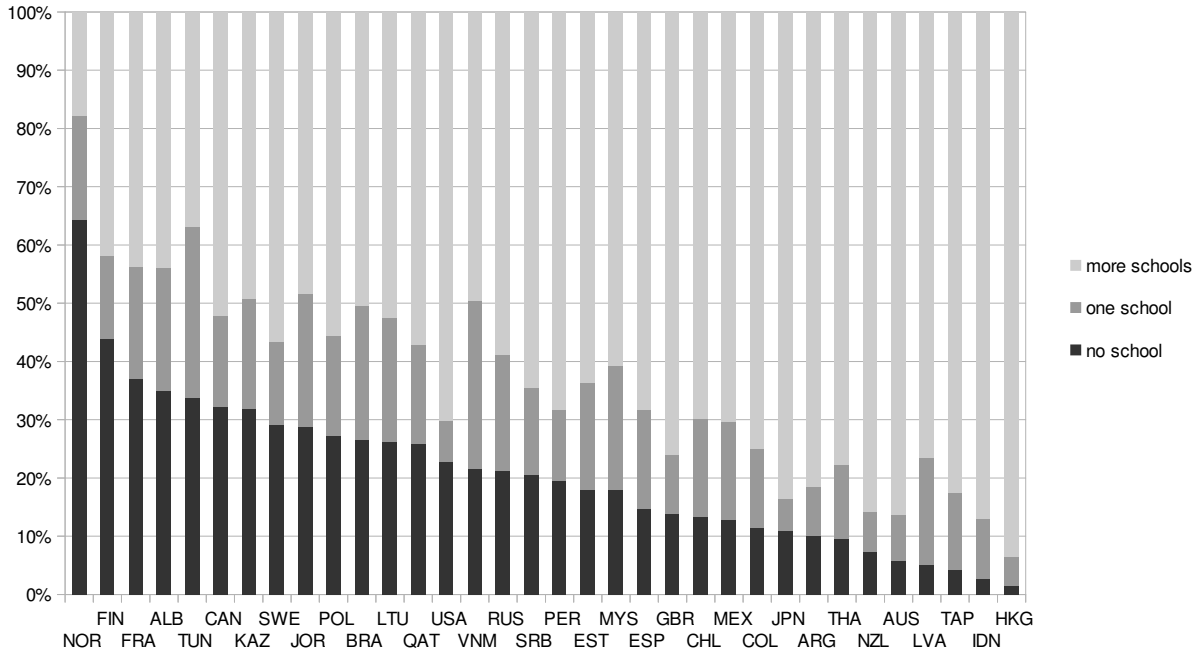

Figure 6: Number of Schools a School Competes with in PISA 2012

Furthermore, we argue that the instrument is valid in terms of it not being correlated with the dependent variable *achievement*. First of all, the fact that there are more or less schools in a region is mostly exogenously given from historic and geographical reasons. One could argue that more schools open in regions where residents' education is high, and students are expected to be of high ability. It can be shown, however, that the correlation between $Schoolcomp_{sc}$ and $Books_{isc}$ (a proxy for students ability and family background) is very low (0.05). In fact *location* has the highest correlation with $Schoolcomp_{sc}$ (0.39), showing that the bigger the town, the more schools compete with each other. This underlines the exogenous character of $Schoolcomp_{sc}$. See also Currie and Moretti (2003) for arguments in favor of exogeneity of the number of schools in a given area. Furthermore, it might be argued that school competition improves a schools' quality, leading to a positive correlation between achievement and $Schoolcomp_{sc}$. However, research shows that there is no significant positive link between active school choice and achievement. These studies use randomized lotteries due to the highly selective nature of students who chose their school (see Musset, 2012, p.25).

In the IV approach the variable AG_{sc} is not used in dummy coding, since more instruments would then be needed.[7] Still, the endogenous variable AG_{sc} appears twice in the main regression. Once on its own and once in the interaction with $Comp_c$. Therefore there are two endogenous variables in the regression for which we need two instruments. According

[7]Using an approach with AG_{sc} in dummy coding and only instrumenting the dummy for $AG = 5$ (all classes grouped) yields roughly the same results as those presented here.

to Wooldridge (2002, pp.121) the natural instrument for an interaction is to substitute the endogenous variable in the interaction with the instrument. Thus, $Schoolcomp_{sc} \times Comp_c$ is the instrument used for $AG_{sc} \times Comp_c$.

Table 6 yields the results of the first-stage regressions from a two-stage-least squares approach as well as for a baseline model that does not include the $Comp_c$ interaction. The results illustrate that $Schoolcomp_{sc}$ is positively correlated with the endogenous variable AG_{sc} in the baseline regression. Once the interaction with $Comp_c$ is included the coefficient on $Schoolcomp_{sc} \times Comp_c$ is positively significant, suggesting that the more competitive a country, the more do students choose ability-grouped schools. This indicates that the more competitive a student's attitude, the more do they actively seek a competitive environment, i.e. ability-grouped schools, if they have the choice. Students in non-competitive countries do not seem to actively choose comprehensive or ability-grouped schools.[8]

Table 6: First-Stage Regressions

Variables Dep. Variable	(1) Ability Grouping	(2) Ability Grouping	(3) Ability Grouping × Comp
Schoolcomp	0.106***	-0.139*	-0.653*
	(0.041)	(0.077)	(0.389)
Schoolcomp × Comp		0.049***	0.283***
		(0.018)	(0.109)
Student controls	Yes	Yes	Yes
School controls	Yes	Yes	Yes
Country FE	Yes	Yes	Yes
Country obs.	34	34	34
School obs.	10,534	10,534	10,534
Student obs.	249,505	249,505	249,505
R^2	0.09	0.10	0.29
Robust Fstat	6.74***	9.52***	9.44***
Hausman	0.42	3.88	

Notes: Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, books at home, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

[8] Appendix D also lists results for the cluster-robust OLS regression on male and female subgroups. Conducting the IV analysis only on male students, shows that F-tests on the first stage are higher, indicating more selective behavior of male students. The OLS analysis on these subgroups, however, shows that male and female students are equally affected by AG.

The cluster-robust F-statistic on the excluded instruments in the baseline specification (1) is too low for an IV approach. Also in the model including the interaction with $Comp_c$ (specification (2) and (3)) the F-statistic is just below 10, the level which is usually recommended as proof of strong instruments. In addition, a Hausman test is conducted, which is a test of the exogeneity of AG_{sc}. The chi-squared statistic for the significant error terms of the first-stage regressions included in the OLS regression is not significant in either model. This indicates that there is no evidence of endogeneity of AG_{sc}. However, the Hausman test is only as good as the instrument used and in case of a weak instrument might fail to diagnose endogeneity correctly (Hahn and Hausman, 2003). To increase the power of the instrument first-stage regressions on different subgroups of the population are conducted (similarly in Figlio and Page, 2002). This might increase the power since the monotonicity of the instrument might not be given. Probably not all types of students, high or low-ability, have equal selection behavior. Theoretical predictions from Thiemann (2016) suggest that high-ability students profit from ability-grouped schools, while low-ability students profit from comprehensive schooling. Likewise different selection behavior is expected from different ability groups. As a proxy for student's ability the variable $Books_{isc}$ is used, which indicates in six categories how many books there are at the home of a student.[9] This variable serves as an indicator of parents' education and socio-economic status and should be highly correlated with student's ability, since ability depends to a high degree on genes as passed on by parents and nurture at home (Plomin et al., 1997). Table 7 shows the first-stage results with the $Comp_c$ interaction for students from the six different categories of the variable $Books_{isc}$. Only results for the regression with AG_{sc} as dependent variable are shown (results for the regression with $AG_{sc} \times Comp_c$ are similar). It can be seen that selection only takes place among students with high or medium ability. Again, we observe that students from competitive countries select into ability-grouped schools. For high-ability students from non-competitive cultures we now also find evidence of selection behavior into comprehensive schools. The coefficient on the interaction $Schoolcomp_{sc} \times Comp_c$ is significant even at the 1% level, suggesting that selection behavior in competitive countries is stronger. The lack of significance for students with low ability can be explained with selection criteria of schools. Bad students might thus not even have a choice between schools, since they are not admitted to certain private schools. Also, parents of students from the two lowest subgroups might lack knowledge about strategic school choice or they lack ambition with respect to their child's education. In addition, since $Books_{isc}$ is a variable that also captures the socio-economic status of

[9]Students answered the question on how many books there are at their home themselves. To illustrate the numbers of the six categories pictures of bookshelves were shown. It was also mentioned that schoolbooks should not be included (OECD, 2012).

Table 7: First-Stage Regressions on Sub-Samples

Variables Sample	(1) Books>500	(2) Books201-500	(3) Books101-200	(4) Books26-100	(5) Books11-25	(6) Books0-10
Schoolcomp	-0.300**	-0.243**	-0.165*	-0.168*	-0.049	-0.070
	(0.123)	(0.116)	(0.094)	(0.086)	(0.075)	(0.105)
Schoolcomp						
× Comp	0.093***	0.077***	0.072***	0.065***	0.019	0.030
	(0.026)	(0.024)	(0.020)	(0.019)	(0.018)	(0.024)
Student contr.	Yes	Yes	Yes	Yes	Yes	Yes
School contr.	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34	34	34
School obs.	5,531	6,909	8,476	9,944	9,564	9,305
Student obs.	13,163	23,374	32,956	67,487	52,877	59,648
Avrg. R^2	0.18	0.17	0.17	0.12	0.08	0.06
Robust Fstat	14.72***	13.57***	26.22***	16.40***	1.25	2.80
Hausman	0.80	0.66	1.06	3.11	5.94*	2.17

Notes: Dependent variable: Ability Grouping. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

parents, they might not have the money to send their students to expensive private schools. The F-statistic for the significance of the instruments is well above ten in the first four subgroups of Table 7, which is a good foundation for an IV regression on these subsamples. Performing Hausman tests of endogeneity for the different subgroups, indicates that we can reject endogeneity of the variable AG_{sc} (see Table 7). None of the Hausman tests in the first four subgroups is significant, suggesting that the OLS estimates are the true estimates. In these four subgroups selection does seem to take place, but either to such a little extent that the OLS estimates are not biased or the inclusion of control variables on student background renders the omitted variable problem non-existent. As for the lowest two subgroups (books 11-25, books 0-10), the first-stage regression suggests that there is no self-selection of students. Therefore, OLS estimates yield the true estimates also for these subgroups. The results from the OLS and the quantile regression in Section 5 can thus be considered as robust to endogeneity and unbiased.

As a robustness check we repeat the IV analysis of Equation (3) without including student level controls on family background[10]. This is done to verify that our variable for grouping would be endogeneous in a regression without any proxy for innate ability. The Hausman test now yields significant results (chi-squared statistic: 8.59**; p-value: 0.0136) indicating that we can reject exogeneity of AG_{sc}. This shows that the inclusion of family

[10]i.e. without *books at home*, *hisei* and *parents' education*

background variables, as done in all our regressions, can proxy effectively for unobserved innate ability such that the Hausman test is insignificant.

7 Further Robustness Checks

Several additional robustness checks are conducted. First, a cluster-robust OLS regression as in Equation (3) is conducted for the dependent variable *science* and *reading* achievement (see Appendix F). The regressions yield roughly the same results as the reported ones in Section 5.1 with math achievement. For science significance is even stronger, for reading less strong. In addition, the OLS analysis is run with different definitions of the AG_{sc} variable. For instance, a question from the school questionnaire on *within-class* grouping can be included to define a variable $AG2_{sc}$ that takes on the following values: 0 if "not for any classes" was ticked both for between-class grouping and for within-class grouping; 1 if "not for any classes" was ticked for between-class grouping; 2 if between-class grouping is operated for "some classes" and 3 if between-class grouping is operated in "all classes". The results for the cluster-robust OLS analysis are given in Appendix F. Again, results are similar to those reported in Section 5.1. Direction and significance of the effects are the same, only the coefficients are a bit smaller in size once $AG2_{sc}$ is considered. Within-class grouping has no significant effects.

For the results presented in the main part of this paper we decided to drop all observation of first and second generation immigrants, since national culture of the test country cannot be assigned to them. Since the effect of competitiveness is assumed to work via peer effects, the immigrant population could still matter in the sense that native students are affected by the performance of the immigrant students in their class. To account for this we repeat the regression as in Equation (3) including the immigrant population, but controlling for their immigrant status, also in an interaction with AG_{sc}. The results as given in Appendix F.3 show that the coefficients on AG_{sc} and $AG_{sc} \times Comp_c$ do not change compared to those reported in Section 5.1. The coefficients on the controls for immigrants are insignificant.

An estimation technique very often used in educational research is multi-level-analysis, i.e. a random effects model that takes into account the different levels of observation of the data, namely student, school and country level. Since the interest of this paper is only in the effects at the school level, so far only the OLS analysis was presented with country fixed effects and standard errors adjusted for the country clusters. However, the results of a random effects model shall be given as a robustness check (see Appendix F). The results are qualitatively the same as the OLS results presented in Table 4, with coefficients on

AG_{sc} and $AG_{sc} \times Comp_c$ being a bit bigger in size. The regression results also show that the model can explain almost 70% of the between-school variation, but only 12% of the within-school variation.

8 Conclusion

The analysis of school level PISA 2012 data has shown that culture, or more precisely competitiveness, *does* matter for the effect of AG on student performance. Particularly, we find evidence for AG being detrimental in non-competitive cultures, but beneficial in competitive cultures. Students at the tails of the ability distribution are generally less affected than those closer to the median. The effect of AG on the variance of achievement is not significantly different from zero.

The positive effect of AG in competitive cultures supports the idea that being surrounded by students of similar ability can be more motivating than being in a class with students of heterogeneous abilities. This positive effect of AG can be explained by the model from Thiemann (2016) including a non-linear value function, thereby modeling diminishing sensitivity to the reference point. For instance the value function of Tversky and Kahneman (1979) is convex below the reference point, indicating that being just below the reference point induces a higher motivation than being further away. Another explanation can be the existence of a participation constraint (Thiemann, 2016). Students that give up because of being too far away from the reference point are mainly a problem in comprehensive schools, where abilities are very heterogeneous. Under AG, however, the reference point is usually close enough to drive students to perform. Furthermore, competitive students in ability-grouped schools might be incentivized by the chance of being promoted to a higher track, if they perform among the best of the group. This possibility has not been considered in the theoretical model and is subject to further research.

In non-competitive cultures evidence is found for students losing under AG, especially medium to low-ability students. The latter coincides with theory and can be explained by students in lower tracks having a lower reference point than under comprehensive schooling. Especially the relatively good students in the low track are not motivated anymore, since they have no-one to look up to. The overall detrimental effects of AG in non-competitive cultures could also be due to some kind of "competition-aversion", which is not covered by the theory of Thiemann (2016). The model includes the possibility that students are non-competitive in the sense that they do not get any utility or disutility from social comparison. Then students' utility increases only in own performance. It might be possible, however, that relative performance feedback has a discouraging effect. For

example students that feel comfortable being mediocre in a comprehensive school, would find themselves being a bad student in a high track of a grouped system. While this might drive competitive students to perform higher it might demotivate non-competitive students, because of too high expectations and pressure to perform. Correspondingly the IV approach has shown that competitive students actively seek more competitive environments (ability-grouped schools), whereas students from non-competitive cultures avoid this.

All in all the analysis has provided an important contribution to the existing literature by showing that there is a significant effect of AG on student performance once we distinguish between competitive and non-competitive cultures. This reveals that school systems have to be designed taking into account the culture in a given country. However, with field data from PISA it is hard to investigate the structure of incentives that drives students to perform at a certain level. For further research a laboratory experiment might be useful to disentangle the channels that drive subjects to perform and test the theoretical hypotheses in an environment that closely matches the model from Thiemann (2016). In an experiment confounding factors can be eliminated and factors considered in theoretical models (loss aversion, individual reference points, competitiveness) can be tested for directly.

References

Ammermüller, A. (2005). Educational opportunities and the role of institutions. *ZEW Discussion Papers 05-44*.

Argys, L. M., D. I. Rees, and D. J. Brewer (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management 15*(4), 623–645.

Betts, J. R. and J. L. Shkolnik (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review 19*(1), 1–15.

Brunello, G. and D. Checchi (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy 22*(52), 781–861.

Cortes, K. E. and J. S. Goodman (2014). Ability-tracking, instructional time, and better pedagogy: The effect of double-dose algebra on student achievement. *The American Economic Review 104*(5), 400–405.

Currie, J. and E. Moretti (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *The Quarterly Journal of Economics 118*(4), 1495–1532.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review 101*(5), 1739–74.

Figlio, D. N. and M. E. Page (2002). School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics 51*(3), 497–514.

Fuchs, T. and L. Wößmann (2008). What accounts for international differences in student performance? A re-examination using PISA data. In C. Dustmann, B. Fitzenberger, and S. Machin (Eds.), *The Economics of Education and Training*, pp. 209–240. Springer.

Galindo-Rueda, F. and A. Vignoles (2007). The heterogeneous effect of selection in UK secondary schools. In L. Woessmann and P. E. Peterson (Eds.), *Schools and the Equal Opportunity Problem*, Chapter 5, pp. 103–128. CESifo Seminar Series.

Guiso, L., P. Sapienza, and L. Zingales (2003). People's opium? Religion and economic attitudes. *Journal of Monetary Economics 50*(1), 225–282.

Hahn, J. and J. Hausman (2003). Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review 93*(2), 118–125.

Hanushek, E. and L. Woessmann (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal 116*(510), C63–C76.

Hanushek, E. A., S. Link, and L. Wößmann (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics 104*(2013), 212–232.

Hanushek, E. A. and L. Woessmann (2014). Institutional structures of the education system and student achievement: A review of cross-country economic research. In R. Strietholt, W. Bos, J.-E. Gustafsson, and M. Rosen (Eds.), *Educational Policy Evaluation through International Comparative Assessments*, pp. 145–175. Waxmann Verlag.

Hayward, R. D. and M. Kemmelmeier (2007). How competition is viewed across cultures. A test of four theories. *Cross-Cultural Research 41*(4), 364–395.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis 14*(3), 205–227.

Hofstede, G. (1984). *Culture's Consequences: International Differences in Work-Related Values*, Volume 5 of *Cross Cultural Research and Methodology Series*. Newbury Park: SAGE Publications.

Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of Intercultural Relations 10*(3), 301–320.

Horton, N. J. and K. P. Kleinman (2007). Much ado about nothing. *The American Statistician 61*(1), 79–90.

Inglehart, R. (2014). World Values Surveys and European Values Surveys, 1981-1984, 1989-1993, 1994-1999, 1999-2004, 2005-2007 and 2010-2014. www.worldvaluessurvey.org.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74–89.

Macdonald, K. (2014). PV: Stata module to perform estimation with plausible values. *Statistical Software Components*.

Manning, A. and J.-S. Pischke (2006). Comprehensive versus selective schooling in England and Wales: What do we know? *IZA Discussion Paper 66*.

Meier, V. and G. Schütz (2007). The economics of tracking and non-tracking. *Ifo Working Paper 50*.

Musset, P. (2012). School choice and equity: Current policies in OECD countries and a literature review. *OECD Education Working Papers 66*.

OECD (2005). *PISA 2003 Data Analysis Manual*. OECD Publishing.

OECD (2012). *OECD Programme for International Student Assessment 2012. Student Questionnaire - Form A*. OECD Publishing.

OECD (2013a). *PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices. (Volume IV)*. OECD Publishing.

OECD (2013b). *PISA 2012 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I)*. OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*. OECD Publishing.

Parente, P. M. and J. Santos Silva (2013). Quantile regression with clustered data. *University of Essex Discussion Paper Series 728*.

Pekkarinen, T., R. Uusitalo, and S. Kerr (2009). School tracking and intergenerational income mobility: Evidence from the Finnish comprehensive school reform. *Journal of Public Economics 93*(7), 965–973.

Plomin, R., D. W. Fulker, R. Corley, and J. C. DeFries (1997). Nature, nurture, and cognitive development from 1 to 16 years: A parent-offspring adoption study. *Psychological Science 8*(6), 442–447.

Schafer, J. L. and J. W. Graham (2002). Missing data: Our view of the state of the art. *Psychological Methods 7*(2), 147.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research 60*(3), 471–499.

The Economist Intelligence Unit (2012). *The learning curve. Lessons in country performance in education.* Pearson.

Thiemann, K. (2016). Ability tracking or comprehensive schooling? A theory on peer effects in competitive and non-competitive cultures. *University of Hamburg*.

Tversky, A. and D. Kahneman (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–291.

VanderHart, P. G. (2006). Why do some schools group by ability? *American Journal of Economics and Sociology 65*(2), 435–462.

Woessmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics 65*(2), 117–170.

Woessmann, L. (2008). How equal are educational opportunities? Family background and student achievement in Europe and the US. *Zeitschrift für Betriebswirtschaft 78*(1), 45–70.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

World Bank (2014a). Expenditure per student, secondary (% of GDP per capita). http://data.worldbank.org/indicator/SE.XPD.SECO.PC.ZS/countries.

World Bank (2014b). GDP per capita (constant 2005 USD). http://data.worldbank.org/indicator/NY.GDP.PCAP.KD.

Appendix

A Measure of Competitiveness from WVS

Country	Code	Competitiveness
Jordan	JOR	9.995
Tunisia	TUN	9.79
Albania	ALB	7.878
Latvia	LVA	7.148
USA	USA	6.980
New Zealand	NZL	6.819
Estonia	EST	6.740
Sweden	SWE	6.681
Australia	AUS	6.653
Indonesia	IDN	6.639
Peru	PER	6.510
Norway	NOR	6.312
Lithuania	LTU	6.228
Qatar	QAT	6.09
Serbia	SRB	5.961
Mexico	MEX	5.648
Canada	CAN	5.338
Brazil	BRA	5.299
Taiwan	TAP	5.222
Vietnam	VNM	5.179
Hong Kong	HKG	5.068
Colombia	COL	4.895
Finland	FIN	4.638
Russia	RUS	4.631
Malaysia	MYS	4.511
Argentina	ARG	3.714
United Kingdom	GBR	3.550
Spain	ESP	3.184
Japan	JPN	2.718
Poland	POL	2.346
Chile	CHL	1.862
Thailand	THA	1.155
Kazakhstan	KAZ	0.8
France	FRA	0.012

Notes: Reverse coded and normalized to values from 0 (non-competitive) to 10 (competitive).

B Missing Values

Including all control variables would result in a loss of almost 60% of the data if observations with missing values are dropped. 40% of the observations have one missing value, more than 19% of the observations even more. There is no pattern of missing values, but the values seem to be missing at random (MAR) in a non-monotone manner. Most values are missing for the variable *classsize*.

Table 8: Summary Statistics

Variable	Mean	Std. Dev.	N
Achievement	447.049	101.297	252,921
Age	15.805	0.292	252,808
Female	0.509	0.5	252,921
Grade Repetition	0.17	0.376	235,239
Other Language at Home	0.135	0.342	248,020
Parents' Education	4.017	1.803	249,324
HISEI	44.479	23.091	235,663
Books at Home	2.645	1.37	248,971
Grade	-0.273	0.727	252,684
Class Size	29.796	9.945	153,976
Number of Students	970.656	795.997	236,018
Private School	0.203	0.402	250,628
Math-Teacher Shortage	0.179	0.383	248,438
Student-Teacher-Ratio	153.548	123.374	224,337
School Location	2.986	1.249	250,378
Government Funding	78.099	33.317	228,977
School Autonomy	-0.048	1.114	250,882
Admission by Ability	2.033	0.894	248,645
Same Textbook	0.737	0.44	245,911

Dropping all students with missing values would result in substantial loss of observations and would lead to biased coefficients. As a solution we impute missing data using the data of students with non-missing data as proposed by Woessmann (2003) and Ammermüller (2005). Unlike using country-by-wave-means for the missing values this does not "distort covariances and intercorrelations between variables" (Schafer and Graham, 2002, p. 159) or introduce bias and understate variability (Horton and Kleinman, 2007, p. 80). Following Woessmann (2003, p.169) the technique works as follows: "For each student i with missing data on a specific variable M, a set of 'fundamental' explanatory variables F with data available for all students is used to impute the missing data. Let S denote the set of students j with available data for M. Using the students in S, the variable M was regressed on F:

$$M_{j \in S} = F_{j \in S}\phi + \epsilon_{j \in S} \tag{4}$$

For M being a discrete variable, OLS estimation was used for the regression. For M being a dichotomous (binary) variable, a probit model was used. If M was originally (before deriving dummies) a polychotomous qualitative variable with multiple categories, an ordered-probit model was estimated. The coefficients ϕ from these regressions and the data on F_i were then used to impute the value of M_i for the students with missing data:

$$\widetilde{M}_{j \notin S} = F_{j \in S}\phi \tag{5}$$

For the probit models, the estimated coefficients were used to forecast the probability of occurrence associated with each category for the students with missing data, and the category with the highest probability was imputed."

As fundamental variables that are complete for almost the whole data set we use student's *age*, *female*, *parents' education*, *wealth*, the *school location*, *GDP per capita* (World Bank, 2014b) and *public spending on education* (World Bank, 2014a). With these fundamental variables values for *grade repetition, other language at home, hisei, books at home, number of students, private school, math-teacher shortage, government funding, student-teacher-ratio, class size, school autonomy, admission by ability, same textbook* and *grade* are imputed. The small amount of missing data within F was imputed by taking the average value at the school level.

C Summary Statistics and Coefficients of Control Variables

Variables	Mean	Std. Dev.	Min.	Max.	Student Obs.
Math achievement	446.698	101.206	19.793	924.84	251,972
Ability Grouping	2.285	1.561	0	5	251,972
Student characteristics:					
Age	15.805	0.292	15.17	16.33	251,919
Female	0.508	0.5	0	1	251,972
Index of Socio-Economic Status (HISEI)	44.173	22.656	-0.652	88.960	250,774
Grade Repetition	0.163	0.35	0	1	251,914
Other Language at Home	0.14	0.343	0	1	251,681
Class Size	29.848	8.137	0	200	251,972
Grade	-0.254	0.700	-3	3	251,972
Parents' education:					
None	0.028	0.165	0	1	251,969
Primary School	0.1	0.3	0	1	251,960
Lower Secondary	0.124	0.33	0	1	251,969
Upper Secondary 1	0.04	0.195	0	1	251,969
Upper Secondary 2	0.411	0.492	0	1	251,897
University	0.296	0.457	0	1	251,969
Books at home:					
Books 0-10	0.259	0.438	0	1	251,835
Books 11-25	0.245	0.43	0	1	251,835
Books 26-100	0.272	0.445	0	1	251,835
Books 101-200	0.111	0.314	0	1	251,835
Books 201-500	0.074	0.262	0	1	251,835
Books $>$ 500	0.039	0.194	0	1	251,972
School characteristics:					
Number of Students	973.801	792.23	1	11483	251,972
Private School	0.203	0.402	0	1	251,972
Math-Teacher Shortage	0.178	0.381	0	1	251,972
Student-Teacher-Ratio	154.462	118.86	0.5	2,311	251,585
Government Funding	78.237	32.171	0	116.302	251,810
School Autonomy	-0.037	1.11	-2.872	1.604	251,972
Admission by Ability	2.028	0.898	0.148	3	251,914
Same Textbook	0.737	0.438	0	1	251,971
School location:					
Village ($<$ 3,000)	0.152	0.359	0	1	251,972
Small Town (3,000-15,000)	0.202	0.401	0	1	251,972
Large Town (15,000-100,000)	0.266	0.442	0	1	251,972
City (100,000-1,000,000)	0.258	0.438	0	1	251,972
Large City ($>$1,000,000)	0.121	0.326	0	1	251,972

	(1)		(2)	
Variables	Coefficients	Std.Error	Coefficients	Std.Error
Ability Grouping	-0.220	0.596	-3.457**	1.441
Ability Grouping × Comp			0.615**	0.292
Student characteristics:				
Age	0.478	1.234	0.465	1.234
Female	-13.900***	0.780	-13.887***	0.782
Index of Socio-Economic Status (HISEI)	0.597***	0.025	0.596***	0.025
Grade Repetition	-33.074***	1.517	-32.994***	1.506
Other language at home	1.276	2.314	1.237	2.298
Class size	0.561***	0.068	0.562***	0.068
Grade	20.715***	1.128	20.765***	1.124
Parents' education:				
Primary school	6.462***	2.015	6.464***	2.016
Lower secondary	5.402***	1.734	5.412***	1.737
Upper secondary 1	6.340**	2.491	6.410***	2.485
Upper secondary 2	6.912***	2.030	6.955***	2.027
University	19.342***	2.093	19.385***	2.090
Books at home:				
Books 11-25	7.696***	0.977	7.664***	0.976
Books 26-100	25.073***	1.133	25.049***	1.134
Books 101-200	36.714***	1.654	36.637***	1.644
Books 201-500	59.923***	1.832	59.930***	1.825
Books > 500	46.294***	2.670	46.285***	2.651
School characteristics:				
Number of student	0.007***	0.001	0.007***	0.001
Private school	-7.080**	3.038	-7.468**	3.012
Math-teacher shortage	-8.201***	2.093	-8.204***	2.085
Student-teacher-ratio	-0.041***	0.009	-0.042***	0.009
Share of government funding	-0.140***	0.036	-0.143***	0.036
Admission by ability	2.383**	1.127	2.373**	1.126
School autonomy	3.734***	1.213	3.800***	1.203
Same textbook	-0.648	2.234	-0.741	2.263
School location:				
Small town (3,000-15,000)	0.257	3.143	0.284	3.131
Large town (15,000-100,000)	0.972	3.162	0.933	3.124
City (100,000-1,000,000)	3.887	3.861	3.954	3.853
Large city (¿1,000,000)	10.343**	4.111	10.418**	4.113
Country FE	Yes		Yes	
Country obs.	34		34	
School obs.	10,558		10,558	
Student obs.	249,968		249,968	
Avrg. R^2	0.49		0.49	

Notes: Dependent variable: PISA math score 2012. OLS regression weighted by students sampling probability. Cluster robust standard errors are given in parentheses. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

D Gender

	Male		Female	
Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.286	-3.454**	-0.098	-3.586**
	(0.690)	(1.741)	(0.644)	(1.467)
Ability Grouping × Comp		0.601*		0.665**
		(0.359)		(0.297)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,074	10,074	10,062	10,062
Student obs.	118,391	118,391	125,344	125,344
Avrg. R^2	0.48	0.48	0.50	0.50

Notes: Dependent variable: PISA math score 2012 of male (female) students. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

E Variance

Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.113	0.120		
	(0.273)	(0.741)		
Ability Grouping × Comp		-0.045		
		(0.145)		
AG=1 (Some classes grouped)			0.611	-0.989
			(1.269)	(3.165)
AG=2			0.744	-0.097
			(1.245)	(2.345)
AG=3			0.596	0.587
			(1.042)	(2.869)
AG=4			-0.222	3.277
			(2.013)	(4.407)
AG=5 (All classes grouped)			-0.400	-1.975
			(1.475)	(3.717)
AG=1 × Comp				0.356
				(0.704)
AG=2 × Comp				0.173
				(0.468)
AG=3 × Comp				0.018
				(0.515)
AG=4 × Comp				-0.622
				(0.865)
AG=5 × Comp				0.301
				(0.774)
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,464	10,464	10,464	10,464
Avrg. R^2	0.34	0.34	0.34	0.34

Notes: Dependent variable: PISA 2012 standard deviation of math test scores per school. Least squares analysis using school weights. Robust standard errors are given in parentheses. Control variables: sd(age), share of females, shares of parents' education, sd(grade), share of grade repeaters, mean class size, sd(books at home), private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math-teachers, school location, admission by ability. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

F Robustness Checks

F.1 Science vs. Reading

	Science		Reading	
Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.020	-3.550***	-0.120	-3.304**
	(0.541)	(1.360)	(0.530)	(1.440)
Ability Grouping × Comp		0.671**		0.605**
		(0.273)		(0.286)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,558	10,558	10,558	10,558
Student obs.	249,968	249,968	249,968	249,968
Avrg. R^2	0.48	0.48	0.44	0.44

Notes: Dependent variable: PISA science score 2012 and PISA reading score 2012. Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

F.2 Alternative Definition of Group

Variables	(1)	(2)	(3)	(4)
Ability Grouping 2	-0.367	-5.052***		
	(0.972)	(1.883)		
Ability Grouping 2 × Comp		0.943**		
		(0.400)		
AG2=2 (Within-Class Grouping)			-4.518	-8.076
			(4.615)	(7.847)
AG2=3 (Some Between-Class Grouping)			-1.760	-18.126***
			(2.942)	(5.185)
AG2=4 (All Classes Grouped)			-1.542	-13.138**
			(3.085)	(5.747)
AG2=2 × Comp				0.785
				(1.893)
AG2=3 × Comp				3.469***
				(1.145)
AG2=4 × Comp				2.491**
				(1.233)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,563	10,563	10,563	10,563
Student obs.	250,042	250,042	250,042	250,042
Avrg. R^2	0.49	0.49	0.49	0.49

Notes: Dependent variable: PISA math score 2012. Reference category is AG2=0 (no within or between class grouping). Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

F.3 Immigrants

Variables	(1)	(2)	(3)	(4)
Ability Grouping	-0.296	-3.709***		
	(0.596)	(1.416)		
First generation	-7.929	-8.018	-7.647	-8.190
	(6.750)	(6.728)	(6.698)	(6.551)
Second generation	-4.288	-3.902	-4.007	-4.677
	(6.786)	(6.795)	(6.749)	(6.751)
Firstgen × Ability Grouping	0.147	0.173	0.079	0.339
	(2.530)	(2.547)	(2.493)	(2.479)
Secgen × Ability Grouping	0.235	0.083	0.144	0.418
	(2.409)	(2.426)	(2.380)	(2.394)
Ability Grouping × Comp		0.649**		
		(0.292)		
AG=1			-1.723	-11.033*
			(3.496)	(6.080)
AG=2			-1.695	-22.305***
			(2.639)	(5.769)
AG=3			0.279	-9.282
			(2.685)	(5.765)
AG=4			-2.151	-19.201**
			(3.258)	(8.193)
AG=5			-2.349	-19.417**
			(3.744)	(8.232)
AG=1 × Comp				2.037
				(1.516)
AG=2 × Comp				4.239***
				(1.209)
AG=3 × Comp				2.100*
				(1.223)
AG=4 × Comp				3.493**
				(1.737)
AG=5 × Comp				3.444**
				(1.656)
Student controls	Yes	Yes	Yes	Yes
School controls	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Country obs.	34	34	34	34
School obs.	10,580	10,580	10,580	10,580
Student obs.	273,720	273,720	273,720	273,720
Avrg. R^2	048	0.48	0.48	0.48

Notes: Dependent variable: PISA math score 2012. Reference category is AG=0 (no between class grouping). Least squares regression weighted by students sampling probability. Robust standard errors adjusted for clustering at the country level are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$

F.4 Multilevel Model

Variables	(1)	(2)
Ability Grouping	-0.347	-4.686***
	(0.579)	(1.567)
Ability Grouping × Comp		0.833**
		(0.324)
Within-school SD	61.61	61.61
Between-school SD	32.47	32.47
Var. prop. attributed to schools (ρ)	0.21	0.22
Within-school var. prop. explained (%)	0.12	0.12
Between-school var. prop. explained (%)	0.70	0.70
Student controls	Yes	Yes
School controls	Yes	Yes
Country FE	Yes	Yes
Country obs.	34	34
School obs.	10,558	10,558
Student obs.	249,968	249,968
Avrg. R^2	0.48	0.48

Notes: Dependent variable: PISA math score 2012. Random effects regression weighted by students sampling probability. Standard errors are given in parentheses. Control variables: Age, female, parents' education, hisei, grade, grade repetition, class size, private school, number of students, government funding, school autonomy, student-teacher-ratio, shortage of math teachers, admission by ability, same textbook, school location. Significance levels: * $p < 0.1$; ** $p < 0.05$; ***$p < 0.01$