LUND UNIVERSITY
School of Economics and Management

Teaching Practices and Collaborative Problem Solving

Evidence from PISA 2015

Jens Mattsson 930522-7036
Email

First year Master thesis

Supervisor: Martin Nordin
Department of Economics
Lund University
2020-08-30

Abstract

There is a heated debate on what teaching practices should be used to foster the skills students need for the current and future workplace. The 21^{st}-century skills movement, among others, argue that 21^{st}-century skills such as creativity, critical thinking, and collaboration should be actively taught to students using more modern teaching practices (for example students working in small groups and focusing on critical thinking) and less traditional teaching practices (for example lecturing and focusing on fact-based knowledge). In this thesis, the relationship between teaching practices and the 21^{st}-century skill collaboration is examined. To investigate this relationship, data on teaching practices and students' Collaborative Problem Solving (CPS) test scores from PISA 2015 was used in OLS regressions including country dummies as well as control variables at the student, teacher, and school level. Results differed significantly based on whether student- or teacher-reported teaching practices were used in the regressions. In the results based on student-reports, traditional teaching practices were indicated to have a large, statistically significant positive effect on CPS test scores, whereas modern practices had a statistically significant negative effect. In contrast, no statistically significant relationships where observed when teacher-reported teaching practices were used. This discrepancy in results based on who reported the teaching practices is discussed and future research is recommended to investigate this further. It is concluded that the relationship between traditional teaching and CPS is either zero or positive, while it is either zero or negative for modern teaching. The policy recommendations made by the 21^{st}-century skills movement in regards to CPS are therefore discredited.

Keywords: Teaching practices, Collaborative Problem Solving, Standardized tests, PISA 2015

Table of Contents

1. Introduction

During the 20^{th} century, one of the most prominent factors deciding wage level was educational attainment, with a high and increasing wage premium ensuring that on average university graduates were paid more than high school graduates (OECD 2017b). However, during the last decade of the 20^{th} century and the first decade of the 21^{st} century, the largest increase in demand was for non-routine analytical skills and non-cognitive (or social) skills that can not so easily be automated by machines or artificial intelligence (OECD 2017b). Thus, the modern and future workplace has been argued to increasingly require proficiency in what has been termed 21^{st}-century skills (Ananiadou and Claro 2009; Echazarra et al. 2016; Kay 2010; OECD 2017b; Rotherham and Willingham 2010). These skills can be succinctly captured by the four Cs: Creativity, Critical thinking, Communication, and Collaboration (Echazarra et al. 2016). This thesis focuses on the last C: students' collaboration ability.

Fostering these 21^{st}-century skills in students during their education is therefore perceived by many as exceedingly important. Education commentators, organizations such as the National Council of Teachers of Mathematics in the USA, and countries like Singapore have in this pursuit called for decreasing "traditional" teaching practices and increasing "modern" teaching practices, though some only argue for an increase in the latter (Echazarra et al. 2016; NCTM 1991; NRC 1996; OECD 2013; OECD 2017b; Zemelman et al. 2005). Modern teaching practices are defined as students working in small groups, relating school material to the real world, and focusing on reasoning and arguing in the pursuit of fostering analytical and critical thought. Traditional teaching practices, in contrast, involve teacher-led lecturing, memorization, practice, and repetition. Modern teaching practices are, then, championed as being better at promoting and developing 21^{st}-century skills, making it vital for policy makers to be able to determine whether this is, in fact, the case.[1]

In order to explore the determinants of the education production function, economists have typically used data sets from large-scale international student assessment tests such as Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Programme for International Student Assessment (PISA), or national standardized tests. Research has been directed at analyzing the effect of factors such as instruction time (Cattaneo et al. 2017; Lavy 2015),

[1] In the short run, the formation of 21^{st}-century skills could affect the individual's chance of going to college and getting a job, but in the long run, these skills are argued to affect the accumulation of knowledge and productivity in the economy. For instance, in most growth models, technology formation, which drives the productivity of an economy, is dependent on education.

class-size (Angrist and Lavy 1999; Wößmann and West 2006) and teaching practices (Bietenbeck 2014; Echazarra et al. 2016; Lavy 2016; Donné et al. 2016; Schwerdt and Wuppermann 2011) on students' achievement on such standardized tests. Generally, in the small literature in economics that has examined the effects of teaching practices on students' mathematics and science test scores, results have differed based on whether teaching practices were student- or teacher-reported. Student-reported traditional teaching practices have typically been found to positively affect student achievement while modern teaching practices have been found to have zero or negative effects (Bietenbeck 2014; Echazarra et al. 2016; Lavy 2016).[2] In contrast, teacher-reported teaching practices have commonly been found to not have any effect on students' test scores (Algan et al. 2013; O'Dwyer et al. 2015).[3] Additionally, teacher-reported teaching practices have been determined to be more reliable and a better proxy for the "true" practices being used (Desimone et al. 2010).

This thesis examines the relationship between teaching practices and students' collaboration ability, in the form of Collaborative Problem Solving (CPS) measured in PISA 2015, in 17 countries. PISA assesses 15-year-old students in mathematics, science, and reading with one or more "special" topic(s) included each time. The 2015 wave is the first (and only to date) to include a part on CPS, and as far as the author knows no other large-scale international student assessment tests have tried to measure this ability.[4] The CPS test attempts to measure how well students can solve different tasks through collaboration. This involves measuring their skills in solving teamwork related problems, such as identifying team members' skills and knowledge, building consensus, and negotiating (OECD 2017b). The definitions of collaboration and CPS are elaborated on further in Section 2.2 and Section 3.4, respectively.

Despite the arguments for the increased relevance of CPS skills and the adoption of government guidelines and recommendations in many countries, the empirical research on what factors that affect CPS is very limited. As Stadler et al. (2019) note, "there is little empirical evidence on how CPS is related to other constructs, its antecedents, or how it can be generally predicted" (p. 6). Previously, Stadler et al. (2019) has investigated the connection between the Big 5 personality types[5] and CPS in PISA 2015 and found that Openness and Agreeableness had a positive effect on CPS ability. OECD (2017b) identified positive relationships between CPS test scores in PISA 2015 and being female,

[2]Despite this general pattern, Bietenbeck (2014), differentiating between the cognitive skills of knowing, applying and reasoning, concluded that modern teaching practices had a positive effect on the reasoning part of the standardized test in TIMSS 2007.

[3]Algan et al. (2013) did, however, identify a positive relationship between modern teaching practices and the non-cognitive measure social capital.

[4]PISA themselves writes that "PISA 2015 is the first large-scale, international assessment that tries to evaluate competency in collaborative problem solving" (OECD 2017b, p. 52).

[5]Extroversion, Agreeableness, Openness, Conscientiousness, and Neuroticism

attending high socio-economic profile schools, feeling safe at school, and being treated fairly by teachers. OECD (2017b) also found a slight negative effect of "communication intensive" (defined as modern) teaching practices on CPS skills. However, they used only a small amount of control variables, and merely looked at the student-reported teaching practices, disregarding the available data on teacher-reported teaching practices. In light of the discrepancy in results regarding students' mathematics and science test scores based on whether student- or teacher-reported teaching practices are used, these results seem even less convincing and conclusive. Accordingly, this thesis will also investigate if there is a difference in the relationship between teaching practices and CPS test scores for student- and teacher-reported teaching practices.

The contribution of this thesis to the emerging literature on this subject is therefore twofold. First, this is the first time teaching practices, defined along the traditional-modern division used in, among others, Algan et al. (2013), Bietenbeck (2014), and Lavy (2016) is related to CPS with the inclusion of such a rich set of control variables. Second, this thesis aims at investigating the possible heterogeneity in the relationship between teaching practices and CPS, based on whether teaching practices are reported by students or by teachers.

An important problem in identifying the relationship between teaching practices and CPS is the bias induced by unobservable variables. For example, students and teachers with higher unobserved ability might self select into schools where modern teaching practices are employed to a greater extent, introducing bias in any naive estimate. The problem can be tackled in different ways with student fixed effects being the most common; instrumental variables estimates are typically not employed in this context since instruments that would not violate the exclusion restriction are difficult to find (Algan et al. 2013).

If each student in the PISA data would have been tested in CPS in two subjects, and the questionnaire had asked questions regarding teaching practices in both subjects, the variation in teaching practices within the student, between the two subjects, could have been exploited. The data would then form a panel with two observations for each student, one in each subject. By using within student fixed effects estimations, all time-invariant student level and school level variables (both observed and unobserved) could have been removed, and implicitly controlled for. Thus, any bias induced by such unobservable variables would be removed.

However, since PISA samples at the school and student level during a certain year, and the questionnaires only ask about teaching practices in one subject (science), the PISA 2015 data does not allow for clever empirical approaches such as student fixed effects. This means that, due to data limitations, the only identification strategy left to employ is Ordinary Least Squares (OLS) with controls. Nevertheless, from the PISA questionnaires a rich set of control variables for student, teacher, and school characteristics can be included in the regressions (compare Caro et al. 2016; Fuchs and Wößmann 2008; O'Dwyer et al. 2015). The inclusion of these controls alleviates some of the probable endogeneity and bias in the estimates caused by self-selection into teaching practices at the student, teacher, and school level. Considering the possible remaining bias, it is not possible to make statements regarding causality. Ultimately, the analysis provided elucidates part of a broader picture, and might together with the previous literature form a starting point for future research. The hope is that the evidence presented in this thesis can act as a benchmark for future research that can find and use data where it is possible to control for unobserved variables.

The results indicated a positive effect of traditional, and a negative effect of modern, teaching practices on CPS when student-reported teaching practices were used. When teacher answers were used, no statistically significant relationship between CPS and teaching practices was observed. These results roughly corroborate those in the previous literature relating teaching practices to mathematics and science test scores.

Since teacher-reports are plausibly more reliable (Desimone et al. 2010), the results based on student-reports imply that student *perception* of what teaching practices are used affect CPS. It is not clear what these perceptions are measuring or what the results based on them mean for policy, since basing education policy not on what teaching practices are truly employed, but on student perception of these is problematic. Therefore, further research on the subject is encouraged, especially with data where it is possible to employ student fixed effects or similar strategies that can alleviate omitted variables bias. A recommendation would be to also conduct smaller, more qualitative studies to determine what the student perceptions of teaching practices actually measure and how they are formed. It is concluded that traditional teaching practices have either a positive or a zero effect on CPS, whereas modern practices have either a negative or a zero effect. Acknowledging the inability to identify causal relationships, these results still discredit the policy recommendations that call for more modern, and less traditional, teaching practices to foster CPS skills. Policy makers are instead recommended to remain cautious in making any large shifts away from traditional teaching practices.

The remainder of this thesis is organized as follows. Section 2 explores the concept of 21^{st}-century skills, how collaboration can be defined, and the differences between student- and teacher-reported teaching practices. Section 3 introduces the PISA data, how teaching practices and CPS are defined and measured, as well as presents descriptive statistics. Section 4 describes the empirical approach. Section 5 presents the empirical results and discusses them. Section 6 concludes.

2. Theoretical background

2.1 21^{st}-century skills

During the majority of the 20^{th} century, students in developed countries were mostly taught through lectures and working routine problems, with a focus on rote memorization. Cognitive skills and general knowledge was what made one desirable in the workplace (OECD 2017b). As more jobs were automated through robots and computers, voices that wanted curriculum reform and changes in how teachers taught students started to be raised. In the US, for example, National Teaching Standards (a collection of teaching recommendations in various subjects crafted by professional education organizations) started advocating for modernizing teaching to better prepare students for the modern world and workplace (see summary on the movement in Bietenbeck 2014). In essence, the argument was that teaching should be shifted "from traditional, teacher-centered teaching towards modern, student-centered teaching in order to promote students' reasoning skills." (Bietenbeck 2014, p. 145).

In the same strain, the phrase 21^{st}-century skills started to be used in the 1980s as an umbrella term for the skills and abilities deemed necessary for students to be proficient at in the future. Since then the term has become almost ubiquitous in the education debate, with proponents purporting different (but similar) definitions of what skills are defined as 21^{st}-century skills. Definitions typically include critical thinking, Information and Communication Technology (ICT) proficiency, communication, collaboration and effective contribution, creativity and innovation, flexibility and adaptability, leadership, and cross-cultural skills (Ananiadou and Claro 2009; Echazarra et al. 2016; Kay 2010; OECD 2017b; Rotherham and Willingham 2010).

Among the biggest proponents for focusing more on 21^{st}-century skills are organizations such as the Organisation for Economic Co-operation and Development (OECD) (e.g. OECD 2013), Partnership for 21st Century Learning (2020) and Common Core Standards (2020). They argue that for students to become proficient in these vital skills, they must be taught to students while at the same time making instruction more student-oriented. Thus, a theoretical connection between the teaching of these skills and students' abilities is implied.

Critics of this movement tend to object to the phrase "21^{st}-century skills", contending that these skills are not new for the 21^{st} century and nothing implies that they should be relevant for the whole century (Kay 2010; Lucas 2019). Some organizations argue that the most important thing for students is to attain fact-based knowledge; for without such knowledge, it is not possible to analyze, contrast, or critically assess (Core Knowledge foundation 2020). Many of the skills referred to as 21^{st}-century skills are argued to develop regardless of teaching practices, and some are not thought possible to teach effectively. Practices such as self-directed learning and shifting responsibility more to the students are regarded as negative to students' learning outcomes. The concern is that too much focus on trying to teach these skills and letting students manage their learning will have a crucially negative effect on students' level of fact-based knowledge. In essence, the argument is that skills like critical thinking or collaboration skills are meaningless without a fact-oriented foundation (Kay 2010; Rotherham and Willingham 2010).

Nonetheless, the 21^{st}-century skills movement has had many breakthroughs in the last 20 years. A prominent example is that most OECD countries have been inspired to change their regulations, guidelines, or recommendations for compulsory education in order to implement the teaching of these skills (Ananiadou and Claro 2009). The changes have lead to more use of modern teaching practices. In the US, specifically, the Partnership for 21st Century Learning has 14 states in collaboration with them (Partnership for 21st Century Skills 2020) and Common Core was adopted by 41 states in 2010 (Common Core Standards 2020).[6] However, according to Ananiadou and Claro (2009), standardized assessment of 21^{st}-century skills are difficult and not common in the OECD countries they examined; one reason being that 21^{st}-century skills are described as "rather ill-defined" (Ananiadou and Claro 2009, p. 14). Another reason is that it is not clear how to measure proficiency in communication or collaboration.

It is in this context that PISA, in the pursuit of developing standardized assessments for these skills, has been developed. The test items in mathematics, science, and reading have been formulated to test the ability of students to use their knowledge and skills to tackle real-life problems rather than testing their ability to solve routine problems. In some waves, PISA has included sub-topics such as financial and ICT literacy; in 2012 assessment of individual, creative problem-solving skills were included, and PISA 2015 assessed Collaborative Problem Solving skills which is the focus of this thesis.

2.2 Collaboration

The 21^{st} century skill focused on in this thesis is collaboration, which has been identified as a particularly important skill (e.g. Stadler et al. 2019; Hughes and Jones 2011). Some sort of collaboration or teamwork is needed in most situations; whether it is at school, at the workplace, or during social activities. In particular, workplaces have been increasingly demanding that prospective employees be proficient at collaboration (Hughes and Jones 2011; OECD 2017b). In this context, the importance of developing one's collaboration ability was described by OECD (2017b) as follows:

[6] Although, since then some of these states have either repealed the adoption or decided not to implement them.

> Collaborative problem solving has several advantages over individual problem solving: labour can be divided among team members; a variety of knowledge, perspectives and experiences can be applied to solve the problem; and team members can stimulate each other, leading to enhanced creativity and a higher quality of the solution. But collaboration also poses potential challenges to team members. Labour might not be divided equitably or efficiently, with team members perhaps working on tasks they are unsuited for or dislike. Conflict may also arise among team members, hindering the development of creative solutions. Collaboration is thus a skill in itself. (OECD 2017b, p. 32)

Hence, being able to collaborate well can have many advantages and is seen as an individual skill. There are many different descriptions of what constitutes collaboration and teamwork, and what areas should be assessed when trying to gauge a persons' aptitude in them.

Drawing from the literature on the subject, Wang et al. (2009) identified five general content areas of collaboration: task-related process skills, cooperation and communication in the group, influencing team members through support and encouragement, conflict resolution via negotiation strategies, as well as guidance and mentor-ship of other team members. Similar definitions are found in Hughes and Jones (2011) and Nelson (1999) with additional areas of interest described as "goal setting and performance management" and "planning and task coordination" (Hughes and Jones 2011, p. 57). It is apparent that also PISA 2015 followed approximately the same definitions described in the literature when assessing CPS, see Section 3.4.

When it comes to how this skill is to be developed in students, the most common position of 21^{st}-century skills proponents is that collaboration skills are tied to instructional theory (Hughes and Jones 2011; Nelson 1999).[7] The theory of how to teach and foster collaborative problem solving skills presented in Nelson (1999) is a good representation of this view. He contends that a collaborative and open climate is needed, where the exchange of ideas and information is encouraged, and "Learners should feel free to voice their opinions, explore new ideas, and try out a variety of approaches in their work" (Nelson 1999, p. 247). The guidelines for teachers include focusing on self-directed learning and acting as facilitators for the students while being flexible and tolerant. They should make use of a wide range of teaching strategies, including small and large group discussions and projects, direct instruction, active learning, and just-in-time instruction (Nelson 1999, pp. 249–251).

The teaching practices argued to develop collaboration skills in students are thus unequivocally a step away from traditional lecturing and routine problem solving. In fact, the recommendations essentially call for an increase in more modern teaching practices.

[7] For general critique against this approach see Section 2.1

Moreover, as noted before, modern teaching practices have become more common during the last couple of decades in many countries in response to the recommendations from various organizations. Consequently, examining whether the theoretical implications suggested by 21^{st}-century skills proponents are valid becomes highly relevant. Despite this there exist few national assessments on collaboration skills and the only large-scale international student assessment test is the CPS test in PISA 2015 (OECD 2017b; Stadler et al. 2019).

There exist three principal types of assessment methods for students' collaboration ability: student self-report, teacher-report, and different kinds of assessment tests (Wang et al. 2009).[8] Teacher-reported ability and aptitude tests generally have the highest validity, although all three have been found to have a high degree of inter-correlation (Wang et al. 2009). Hughes and Jones (2011) note that it is important not to equate collaboration skills with how well the end product of a group report or project turned out. In these cases there exists an obvious risk of free-riding, suggesting that to assess collaboration skills effectively, individual tests are needed.[9] The CPS assessment in PISA 2015 is an example of one such approach that is elaborated on in Section 3.4.

It is the above theoretical background and framework, together with past empirical research presented in the introduction, that this thesis is based on. The goal is to examine the relationship between teaching practices and collaboration ability, trying to determine whether the 21^{st}-century skills movement has empirical support for their theoretical claims.

2.3 Student- and teacher-reported teaching practices

One of the objectives of this thesis is to investigate the possible heterogeneity of the relationship between teaching practices and CPS, based on whether teaching practices are reported by teachers or by students. Since education policy often targets what teaching practices should be used, as described above, measuring teaching practices becomes key in evaluating the effects of such policies. There exists a "true", unknown, measure of what teaching practices are employed in classrooms, which is estimated in PISA with the student and teacher answers to questionnaire questions. It is, therefore, crucial to determine whether student- and teacher-reported teaching practices are expected to be the same, and, if they are not, which one is more reliable.

It is not uncommon in the literature that the use of either student-reported or teacher-

[8] It should also be possible to conduct observational studies, where examiners from outside the school observe students collaborating.

[9] However, it is questionable if individual tests can adequately measure the full scope of what constitutes collaboration ability. Nonetheless, such tests should give an indication of the full collaboration ability.

reported teaching practices is simply stated as a matter of fact, and not dwelled upon further.[10] That is even true when both measures are available in the data set at hand, for example in TIMSS 2007 that Bietenbeck (2014) and Algan et al. (2013) used. Even if there is limited available evidence on the subject (see the review of the literature in Desimone et al. 2010), it should be addressed and discussed.

There exist roughly four methods to measure classroom level variables, which come with both advantages and disadvantages: surveys/questionnaires, observation studies, interviews, and experience sampling (Anderson 2019, pp. 161–163). Although observation studies tend to have higher reliability than questionnaires, observation studies, as well as interviews and experience sampling, are very time consuming and expensive to perform. This is the reason why most available data on what teaching practices are used in classrooms come from either student or teacher questionnaires. Both of these measures could be prone to participation bias and, depending on whether the measure concerns the teacher or the student, both forms risk self-report bias.

There exists some research comparing the correspondence between teacher- and student-reported measures such as goal structures in classrooms, teaching practices, teacher-student relationship, motivation, social engagement, and self-regulation, (ACT 2013; Buckley and Krachman 2016; Desimone et al. 2010; Urdan 2010; Wentzel and Muenks 2016). Correlations between the two measures in these studies are generally low or moderate, typically ranging between 0.3 and 0.5 (ACT 2013; Buckley and Krachman 2016; Desimone et al. 2010), and are virtually the same for grades 6 through 9 (ACT 2013). The differences in average mean and standard deviation of teacher- and student-reported teaching practices have been found to be small but statistically significant (Desimone et al. 2010).[11] However, if the correlations are low, and this is caused by systematic, and not random, measurement errors, the two measures will still typically give substantially different results when related to student outcomes (Desimone et al. 2010).

There might exist positive response bias for teacher-reports on modern teaching practices (over-reporting) (Desimone et al. 2010), as teachers are in many countries expected to implement more of these activities. Despite this fear, when reviewing the available literature in a meta-analysis, Desimone et al. (2010) found that teachers' self-reported teaching practices corresponded with both classroom observations and teacher logs. They conclude that teacher self-reported teaching practices are "quite valid and reliable in mea-

[10] See papers referenced in the introduction.

[11] In the study by Desimone et al. (2010) on teacher practices in mathematics, heterogeneity was observed in how similar the student-reported teaching practices were to the teacher-reported ones. Student responses were more similar to the teachers' for the following factors: being female, valuing doing well in mathematics, coming from higher educated homes, having higher test scores in mathematics, being in advanced classes, and being in higher-achieving classes. Nevertheless, although being statistically significant, the differences were small.

suring their instruction" (Desimone et al. 2010, p. 270). They also note that there is little evidence indicating that student reports of teaching practices are a valid measure of actual teachers' instruction and that one should exercise caution in using student reports (Desimone et al. 2010). Furthermore, teachers are the ones who plan lessons and manages the overall instruction. Therefore, it is reasonable to assume that their reports on what teaching practices are used will be a more reliable indicator of the "true" teaching practices employed. Measurement errors from the true values are therefore expected to be larger in student-reported observable classroom measures such as teaching practices, length of lessons, number of tests administered per semester, etc. It is also probable that students who are not fond of the teacher or the subject give substantially less reliable answers than the teacher. Further, it could be the case that students to a greater degree "grade" their teachers when answering questions on objective measures such as teaching practices.

Teacher-reports might be more reliable and have higher validity. However, when it comes to predictive power for student test scores, the picture is more unclear. If student-reported measures simply have higher random measurement errors, the consequence should be attenuation bias. However, in some studies, both measures were correlated with student outcomes, with those reported by teachers having the highest predictive power — see for example Buckley and Krachman (2016) that examined student- and teacher-reported mindsets, essential skills, and habits (MESH) in relation to student outcomes. Elsewhere, student-reported measures had higher predictive power, or only one of the two measures was correlated with student outcomes (see review in Desimone et al. 2010). There is also the mentioned difference in observed predictive power when comparing the results in Bietenbeck (2014) with those in Algan et al. (2013) and O'Dwyer et al. (2015), that used the same data set, TIMSS 2007. Bietenbeck (2014) used student-reported teaching practices, whereas Algan et al. (2013) and O'Dwyer et al. (2015) used teacher-reported teaching practices.

Taken together with the discussion above, if teacher-reported teaching practices are not correlated with an outcome variable, but student-reported practices are, then teaching practices are unlikely to be related to that outcome variable, and vice versa. However, if the student-reported teaching practices are correlated to the outcome variable, this implies that the *perception* of what teaching practices are used is related to the outcome. The use of student-reported perceptions have therefore been argued to be more appropriate when there is no way to gain an objective assessment of a measure, or when the student perception is the desired measure (Buckley and Krachman 2016); for example, students' non-observable mindsets and skills or the students' perceived class environment.

In conclusion, teacher-reported teaching practices can be argued to be more reliable and have higher validity, and student-reports tend to differ significantly from these. Furthermore, results based on student-reported teaching practices should be interpreted more carefully. However, both measures are theoretically relevant and including both in the analysis can give a more comprehensive picture of the relationship between teaching practices and CPS.

3. Data

Data from the 2015 wave of PISA was used in the empirical analysis. PISA is a triennial large-scale international student assessment test, conducted by OECD for both member and non-member partner countries, with the aim to assess 15-year-old students' knowledge and application skills in science, mathematics and reading[12]. Each PISA wave focuses more extensively on one of the three main subjects, with both the first wave performed in 2000 and the most recent one in 2018 focusing on reading.

The 2015 wave covered 540 000 students in 72 countries, this time focusing on science, with reading, mathematics, and CPS as minor areas of assessment. It was the first time that all tests were administered on computers instead of on paper, and it was also the first time that PISA tested student's collaboration ability in the CPS test framework. To the best of the author's knowledge, this is the only large-scale international student assessment in which this was attempted. Thus, despite the limitations of the data, the study is restricted to using the PISA 2015 data in examining the relationship between teaching practices and collaboration skills.

In addition to the assessment tests, students also answered a questionnaire covering a rich set of questions regarding the students themselves, their home environment, and their school and learning experiences (OECD 2017a). Furthermore, principals answered a questionnaire on the school system and the overall teaching and learning environment at their school. Additionally, for the first time in the 2015 wave, an optional teacher questionnaire, including questions on teachers' training, experience, and instructional activities, was offered to schools[13], although only 18 countries participated and provided these data.

[12]Students included were between 15 years 3 months and 16 years 2 months at the time of assessment (OECD 2017b, p. 20) however, for brevity, in this thesis they are referred to as "15-year-old" students, "15-year-old's" etc.

[13]These questions were partially based on the Teaching and Learning International Survey (TALIS), which is another study administered by OECD, geared towards teachers.

3.1 The design of PISA 2015

Similar to other large-scale international student assessment tests, PISA uses specific sampling methods to create representative samples within each country, as well as ensuring validity in comparisons between countries. The sampling method used was the two-stage stratified sampling (OECD 2017b; OECD 2009) where the sampling units were schools and students, instead of schools and classes as is the case in TIMSS. In the first stage, a minimum of 150 schools per country were randomly selected[14], with probabilities proportional to the number of eligible 15-year-old students in each school. In the second stage, 42 students within each sampled school were chosen randomly, with each student having an equal probability of being selected.[15] To account for this sampling design, student sampling weights and replicates included in the PISA data were used in the empirical analysis to ensure unbiased estimates and standard errors (OECD 2009).

In the 18 countries that had the teacher questionnaire option, teachers were randomly chosen within schools in approximately the same manner as for students. Eligible teachers where those who taught the national modal grade for 15-year-old students. 25 teachers per school were sampled, of which 10 were science teachers (OECD 2017c, p. 86).[16]

In PISA 2015, assessments in science, mathematics, reading, and CPS were based on two-hour tests for each student. However, every student did not answer questions in all these subjects. Instead, PISA used a booklet/cluster design where each subject has a number of these booklets containing questions on different areas within that subject. As the main focus of PISA 2015 was science, all students completed two randomly chosen 30-minute booklets in science and two randomly chosen 30-minute booklets distributed among reading, mathematics, and CPS. The design thus implies two things: All students did not answer booklets in all subjects, and for the subjects in which students got booklets, they only answered a small proportion of all possible questions within that subject. Thereby every student was not able to answer, equally thoroughly, questions covering all topics within a subject. That is why PISA used Item Response Theory models (Chapter 5 in OECD 2009; Chapter 15 in OECD 2017c; Wu 2005), where student, school, and country characteristics, as well as the students' own and his or her peers' test scores, were taken into account, to impute so-called "plausible values" (PV). For each subject, stu-

[14]However, in smaller countries like Iceland all schools were included (in this case 124), and some countries "oversampled"; for example Canada that participated with 759 schools. Countries that oversample does so in order to be able to compare results between regions or between ethnic groups within the country (Jerrim et al. 2017, p. 53).

[15]Once again, similar to the first stage, in schools with fewer than 42 eligible 15-year-old's, everyone was included in the sample. Schools with fewer than 20 eligible students could not participate.

[16]In Section 5, the teaching practice indices used in the regressions were constructed from the science teachers' answers on the relevant questions in the teacher questionnaire, while school level teacher control variables were averages from all teachers at that school.

dents received 10 PV that can be defined as random values drawn from a distribution of proficiency estimates, representing “the range of abilities that a student might reasonably have, given the student’s item responses” (Wu 2005, p. 115).[17] Test scores in all subjects tested in PISA were then scaled across all OECD countries to have a mean of 500 and a standard deviation of 100. In the empirical analysis, presented in Section 5, CPS test scores were standardized to have a mean of 0 and a standard deviation of 1 in the full sample (compare Bietenbeck 2014; Lavy 2016).

3.2 Sample selection

In PISA 2015, only 51 countries (410,959 students, in 14,380 schools) out of the 72 included in the whole study had the CPS option, and only 17 of these countries answered the teacher questionnaire.[18] Thus, to be able to compare results based on whether teaching practices were student- or teacher-reported, the full sample in the student questionnaire data set consisted of 155,376 students in 5,483 schools, from 17 countries. In the teacher questionnaire data set, the corresponding full sample consisted of 106,465 teachers in 4,988 schools.[19]

When constructing the teaching practice indices (see Section 3.3), choosing control variables, and merging all data sets, several steps were performed where students and teachers were dropped from this full sample (compare Bietenbeck 2014; Caro et al. 2016). For example, schools with few students or teachers, or where only some students or teachers answered the teaching practice questions, were removed. See Appendix I for a description of all the steps. After performing these procedures, the final sample used in the empirical analysis consisted of 119,702 students and 86,696 teachers (of which 26,336 were science teachers) in 3,714 schools from 17 countries, which means that ca. 23 % of students from the full sample were excluded.[20][21] See Table 3.1 for descriptive statistics in the final sample, divided by country.

[17] This design further inhibits strong claims based on statistical analysis at the student level. Any results from such an analysis should be corroborated by other research.

[18] The 17 countries included in the empirical analysis were: Australia, Brazil, Chile, Colombia, The Czech Republic, Germany, Spain, Hong Kong, Italy, South Korea, Peru, Portugal, China-Taipei, United Arab Emirates, The United States of America, Macau, and the Chinese group of cities and regions named Beijing-Shanghai-Jiangsu-Guangdong.

[19] Of these, the number of science teachers was 27,955.

[20] However, all results reported in Section 5 were robust to all these sample restrictions, see Section 5.5, and are available upon request.

[21] The sample with no restrictions, except limiting the students to the ones in schools that had both students and teachers that answered the questions on teaching practices, was 142,235 students in 4,707 schools.

Table 3.1 Descriptive statistics for the 17 countries in the final sample.

Country	CPS test score	Students	Teachers	Schools
South Korea	547	4,387	2,775	125
Hong Kong	541	5,252	3,154	135
Australia	535	10,928	12,733	544
Macau	535	4,434	2,759	40
Colombia	534	7,724	4,172	210
Germany	534	5,352	7,026	196
China-Taipei	533	6,810	4,307	186
USA	525	5,098	3,391	147
Portugal	508	5,701	3,939	161
Czech Republic	505	5,947	4,854	241
Beijing-Shanghai-Jiangsu-Guangdong	497	9,590	6,254	257
Spain	497	6,247	3,968	179
Italy	483	8,883	7,500	325
Chile	466	5,289	2,910	144
United Arab Emirates	440	11,778	6,941	337
Peru	428	4,434	2,751	136
Brazil	421	11,848	7,262	351
		119,702	86,696	3,714

Notes: Countries are sorted by CPS score. Large dissimilarities in number of students, schools, and teachers between countries are in general because of sampling methods described in Section 3.1. The number of teachers reported are the total number of teachers at each school. The number of science teachers in the 17 countries were 26,336.

In the regressions reported in Section 5, a rich set of control variables from the student, teacher, and principal questionnaires were included. Most of these were variables that other studies have included when trying to explain student achievement in mathematics or science (e.g. Algan et al. 2013; Fuchs and Wößmann 2008; O'Dwyer et al. 2015), but also factors that OECD (2017b) identified as important for CPS abilities in students.[22] In many cases, the included control variables were indices for measures such as socio-economic status or shortage of education resources at the school, created by PISA based on several questions in the questionnaire.[23] For some qualitative categorical variables, such as immigration status and school type, the raw numbers were transformed into dummy variables. Table A11 in Appendix VI reports descriptive statistics for the full list of control variables as well as information on the percentage of missing values for each variable.

[22]Examples of controls at the student level are sex, age, grade, mother's and father's education, immigration status, and learning time in school per week; teacher level examples are employment status, age, sex, teaching experience, and the highest level of education; school level examples are school size in number of students, average class size, school ownership, and student-teacher ratio.

[23]See OECD (2017c) for details on the creation of all indices.

The reason for including information on missing data was that after combining all data sets into one there was an apparent problem regarding missing values for the control variables. Most variables had a low amount while some had ca. 20–30 % missing values (range between 0 and 35.84 %). Given that most students had missing values on some variables, excluding all observations with a missing value on at least one variable would drastically lower the total number of observations[24]. Instead, following Bietenbeck (2014), Fuchs and Wößmann (2008) and Lavy (2016), to make use of all available information all missing values was set to a constant, in this case zero, and included in the regressions dummy variables for missing values for each control variable. If values are not missing at random, running regressions without the missing value dummy variables will produce different results. Such a sensitivity test was, therefore, conducted which is reported in Section 5.5.

3.3 Measuring Teaching practices

Previous studies similar to this thesis have created different measures of teaching practices based on the answers to student questionnaires (Bietenbeck 2014; Bill and Melinda Gates Foundation 2010; Caro et al. 2016; Echazarra et al. 2016; Lavy 2016) or teacher questionnaires (Algan et al. 2013; Donné et al. 2016; O'Dwyer et al. 2015; Schwerdt and Wuppermann 2011); these measures are then related to different outcomes. The ideal would be to have independent observational evaluations on what teaching practices teachers employ (Caro et al. 2016) since students', and to some degree teachers', self-reported answers are likely to be biased (O'Dwyer et al. 2015). Nevertheless, conducting such evaluations would be highly costly and impractical, which results in their unavailability and the common use of self-reported measures of teaching practices in the literature. And, as mentioned in Section 2.3, teacher-reported teaching practices have been found to be reliable and have high validity in estimating the true teaching practices employed. Since PISA 2015 for the first time included both a student and a teacher questionnaire, teaching practices based on both were used in the empirical analysis presented in Section 5.

The most common taxonomy in the economics literature in regards to teaching practices is to differentiate between practices that are more "traditional" or more "modern"[25] (Algan et al. 2013; Bietenbeck 2014; Caro et al. 2016; Echazarra et al. 2016; Lavy 2016;

[24]Only 8.3 % of the observations in the final sample of 119,702 had no missing values on all control variables.

[25]Some alternative ways to define teaching practices are: active learning, cognitive activation and teacher-directed instruction (Donné et al. 2016) or teacher-directed instruction, enquiry-based instruction, perceived feedback and adaptive instruction (Lau and Lam 2017). Nonetheless, in this thesis the terminology and taxonomy employed in Bietenbeck (2014), Lavy (2016) and Schwerdt and Wuppermann (2011) is adopted.

Donné et al. 2016; Schwerdt and Wuppermann 2011). Recall that traditional practices include things such as teacher-led lecturing, memorization, practice, and repetition while modern practices comprise students working in small groups, relating school material to the real world, and focusing on reasoning and arguing in the pursuit of fostering analytical and critical thought. Figure 3.1 succinctly illustrates the division of teaching practices along a Traditional-Modern scale.

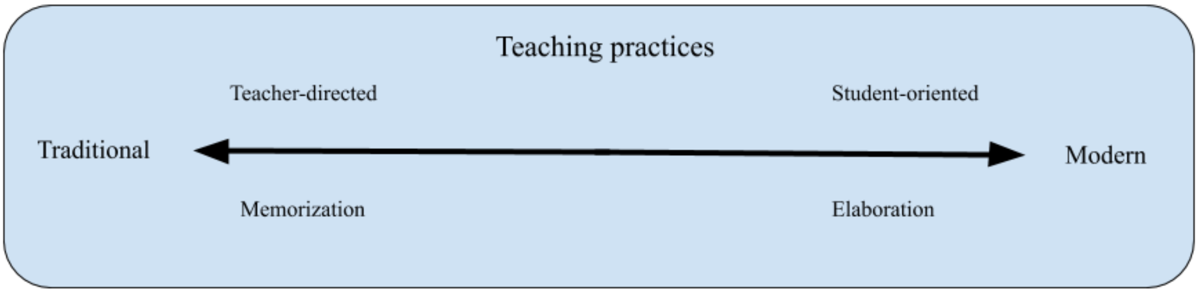

Figure 3.1 Conceptual framework for teaching practices (based on Figure 1.1 in Echazarra et al. 2016)

In the student and teacher questionnaires, the question pertaining to different teaching practices were on the form "When learning <school science> topics at school, how often do the following activities occur?" and "How often do these things happen in your <school science> lessons?" for students and teachers, respectively[26]. These questions were then followed by the different activities: "I explain scientific ideas", "Current scientific issues are discussed", "Students are asked to do an investigation to test ideas" etc. Similar to Bietenbeck (2014) and Lavy (2016), in order to be able to interpret the responses as "the percentage of lessons in which a particular activity was used" (Bietenbeck 2014, p. 146), values on a scale from 0 to 1 were assigned for the different answers in the following way[27]: 0 to "Never or almost never", 1/3 to "Some lessons", 2/3 to "Many lessons" and 1 to "Every lesson or almost every lesson" [28].

Following Bietenbeck (2014), the selection of questionnaire items to assign to either traditional or modern teaching practices was based on the National Teaching Standards from the US (see NCTM 1991; NRC 1996; Zemelman et al. 2005). The teaching practice questions in the teacher and student questionnaires were not the same and the number of questions that could be assigned to either teaching practice was not the same either.

[26]The use of the phrase <school science> is because the same questions are used when mathematics is the focus subject and "science" is then replaced with "mathematics". Also, countries, and even schools within a country, have different names for "science class".

[27]Lavy (2016) found no difference in the evidence produced from the estimates using cardinal, ordinal or categorical values on the questions he included in his teaching practice indices. The cardinal scale was, therefore, used as it is the most intuitive and produces meaningful interpretations.

[28]This was the possible answers in all questions except the questions in the modern index in the student questionnaire. For those questions, the possible answers were somewhat different, but qualitatively the same: "Never or hardly ever ", "In some lessons", "In most lessons", and "In all lessons".

Thus, to enable comparisons between the student and teacher answers, the included questions were three traditional, and one modern, teaching practice that were the same for both questionnaires and two modern teaching practices that, in both questionnaires, qualitatively covered the same sort of activities (see Table 3.2 for the exact wording in the different questionnaires).[29] The traditional teaching practices were when the teacher discusses students' questions, explaining scientific ideas and demonstrating ideas (in essence lecturing), and the modern teaching practices were when students explain their ideas, debate/discuss in groups and work on investigations/scientific research.

Table 3.2 Traditional and modern teaching practice questions from the PISA 2015 student and teacher questionnaires.

Traditional teaching practices	**Modern teaching practices**
Student questionnaire	
The teacher discusses our questions.	Students are given opportunities to explain their ideas.
The teacher explains scientific ideas.	There is a class debate about investigations.
The teacher demonstrates an idea.	Students are asked to do an investigation to test ideas.
Teacher questionnaire	
I discuss questions that students ask.	Students are given opportunities to explain their ideas.
I explain scientific ideas.	A small group discussion between students takes place.
I demonstrate an idea.	Students do their own scientific study\related research.

Notes: Categorization into traditional and modern teaching practices were based on National Teaching Standards (NCTM 1991; NRC 1996; Zemelman et al. 2005).

To gain precision in the estimates and mitigate measurement errors from individual answers, it is common practice in the literature to create teaching practice indices from individual questions and aggregate these to the class-by-school, grade-by-school or school level (Bietenbeck 2014; Blazar 2015; Caro et al. 2016; Lavy 2016; Schwerdt and Wuppermann 2011). As mentioned, PISA 2015 randomly sampled students and teachers in each school and therefore the aggregation was made at the school level.[30] Accordingly, teachers at the same school are implicitly assumed to adopt similar teaching practices which are then contrasted against other schools' teaching practices. This is tantamount to asserting that schools have particular "teaching cultures" where teachers from the same school "collaborate, talk and discuss their teaching strategies" (Donné et al. 2016, p. 9), sharing their teaching practices.[31] Indeed, there is some evidence that this is occurring in schools,

[29] This choice seems unlikely to affect the results in a significant way, see discussion of results in Section 5.3.

[30] The reason is that teachers can not be connected to individual students or classes, only to schools. This, however, allows for the comparison of the relationship to CPS between student- and teacher-reported teaching practices.

[31] See Section 4 for further discussions on what consequences this has for the regression results.

with substantial variation in teaching practices observed between schools (Echazarra et al. 2016, p. 15).

The aggregated indices were created by first calculating the school level mean of the six individual teaching practices and then averaging across the three traditional and the three modern teaching practices. The final school level indices measured the percentage of lessons in which teachers at a school employed the included teaching practices. It is expected that most teachers (in some way) combine the two teaching practices during lessons, implying that there is no "pure" approach (Caro et al. 2016). Also, note that there is not necessarily a one-to-one trade-off between using traditional and modern teaching practices: for instance, a teacher might either spend the whole lesson lecturing, or spend the first half lecturing on a new topic that is then discussed by students in smaller groups, or used as the basis for group projects in the same lesson. Accordingly, a school might have a high value on both indices, and if the modern teaching index were to increase by x, this does not automatically imply that the traditional teaching index decreases by the same amount. There is a time trade-off between different practices during a lesson, but not (within reasonable bounds) a trade-off between using one or many different practices. Indeed, correlations between the traditional and modern indices were 0.46 for student-reports and 0.49 for teacher-reports.

In Table 3.3 the school level mean and standard deviation of the indices is shown. The table also reports the mean and standard deviation across all students/teachers for the questions the indices were constructed from, and the distribution of students' and teachers' answers across the 4 possible answers. The mean of the traditional and modern indices based on student-reports (teacher-reports) were 0.56 and 0.44 (0.67 and 0.54), respectively, with all standard deviations ca. 0.11. Although both indices had a higher mean when reported by teachers, the standard deviations were very similar. For the individual questions making up the indices, means were generally higher when reported by teachers, however, standard deviations were lower. There were also notable differences in the distribution of students' and teachers' answers. Teachers reported more common usage of all traditional teaching practices and, notably, a much higher frequency of letting students explain their ideas. In summary, these statistics indicate that teachers in schools employ both modern and traditional practices in their lessons, and give an early indication of the probable difference in the results based on who reported the teaching practices.[32]

[32] These statistics are in line with the patterns observed in previous research presented in Section 2.3.

Table 3.3 Descriptive statistics for the traditional and modern teaching practice indices.

	Mean	SD	Distribution of students'/teachers' answers (%)			
			Never or almost never	Some lessons	Many lessons	Every lesson or almost every lesson
Panel A: Student-reported indices						
Traditional teaching index	0.56	0.11				
Teacher discusses students' questions	0.54	0.12	12.39	34.15	32.54	20.93
Teacher explains scientific ideas	0.56	0.12	10.15	34.34	30.92	24.59
Teacher demonstrates an idea	0.56	0.11	10.95	32.28	33.99	22.78
Modern teaching index	0.44	0.11				
Students explain their ideas	0.65	0.12	32.83	33.86	25.82	7.49
Class debate about investigations	0.32	0.12	8.65	15.38	36.51	39.46
Students investigate to test ideas	0.38	0.14	10.91	19.50	39.18	30.40
Panel B: Teacher-reported indices						
Traditional teaching index	0.67	0.11				
Teacher discusses students' questions	0.73	0.14	0.82	17.80	44.35	37.03
Teacher explains scientific ideas	0.71	0.15	0.83	21.92	44.77	32.48
Teacher demonstrates an idea	0.58	0.13	3.04	36.54	45.94	14.48
Modern teaching index	0.54	0.11				
Students explain their ideas	0.70	0.13	0.76	20.08	48.19	30.97
Small group discussion between students	0.51	0.16	5.79	47.40	34.34	12.47
Students do their own scientific study	0.37	0.14	19.52	55.51	20.47	4.50

Notes: Means and standard deviations are at the school level for the full sample of 119,702 students, 26,336 teachers in 3714 schools. The possible answers were the same in all questions except the questions in the modern index in the student questionnaire. For those questions the possible answers were "Never or hardly ever ", "In some lessons", "In most lessons", and "In all lessons".

3.4 Measuring Collaborative Problem Solving

PISA 2015 measured students' collaborate problem solving skills using their CPS testing framework, where competency in CPS was defined as:

> the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution. (OECD 2017b, p. 47)

Notice the similarity between this definition (as well as the twelve CPS skills presented below in Table 3.4) and the ones presented in Section 2.2 by for instance Nelson (1999) and Wang et al. (2009). This correspondence implies that the theoretical connection between modern teaching practices and collaboration skills, made by e.g Nelson (1999) and the 21^{st}-century skills movement, should also be expected to hold for CPS skills.

PISA 2015 identified four individual problem-solving processes and three collaborative problem-solving competencies that form a matrix with twelve specific skills (OECD 2017b, pp. 49–50). The individual problem-solving processes were: exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting. Whereas the collaborative problem-solving competencies were: establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organization. Table 3.4 displays the CPS skill matrix where, as an example, the CPS skill "Discovering perspectives and abilities of team members" is a combination of "Exploring and understanding" and "Establishing and maintaining a shared understanding". The twelve skills are what is measured in the CPS items that students have to solve. However, because of small sample sizes in each country, it was not possible to estimate scores in each of the twelve skills and, instead, each student received one final score in CPS.[33]

The correlation coefficient between CPS and mathematics test scores was 0.75, between CPS and science 0.82, and between mathematics and science 0.89. The correlations with CPS are thus quite large, though still lower than between mathematics and science, and could, therefore, indicate that CPS measures something other than general cognitive skills. Another interpretation is that these are preliminary results that indicate that CPS

[33] This score came in 10 PV that incorporated and weighted answers to all the items included in the 30-minute booklets (OECD 2017b, p. 52). The CPS booklets included one or more of jigsaw or hidden-profile tasks (identify who knows what and who has different skills), consensus-building tasks (considering and incorporating the views, opinions, and arguments of all group members to arrive at a group decision) and negotiation tasks (arguing and arriving at a compromise when group members have different opinions) (OECD 2017b, p. 51). See Figure A1 in Appendix II for an example question from one of the CPS units PISA developed: "Xandar". For further information regarding Xandar, see OECD (2017b, Chapter 3) and https://www.oecd.org/pisa/test/.

skills are to a large extent co-determined by those factors that decide the cognitive skills needed for mathematics and science. Regardless, these correlations raise the question of whether CPS manages to isolate the collaboration skill. Simply put, cognitive skills are needed in order to solve the problems presented to the students, although PISA intentionally reduced the difficulty in that regard, so that CPS items only required a low or intermediate ability in individual problem solving, to better isolate the collaboration skill (OECD 2017b, p. 51). Consequently, in the lower part of the cognitive ability distribution, there is a risk that CPS measures, to a significant part, cognitive and not collaborative skills. This concern is explored in Section 5.4.

As mentioned at the beginning of Section 3, students were examined on computers in all test subjects, including CPS where students interacted with programmed computer agents (a minimum of one and a maximum of three for different test items). This raises two questions: 1. Is it meaningful to measure students' CPS abilities in an electronic online setting? 2. Can programmed computer agents in a sufficient way proxy for electronic interactions with humans? Regarding the first question, the workplace of today and that of the future increasingly relies on electronic communication via email, chat programs. etc., making the skill of being able to communicate and collaborate electronically ever more desired by future employers.[34] PISA therefore, in a sense, tests students in realistic future environments. Consequently, assessing CPS skills electronically gives information on a valuable skill for students to possess. It has also been shown that the CPS assessment is informative of how well students can collaborate in real-life scenarios with other humans (OECD 2017b, p. 49). The second question was examined by Herborn et al. (2020) who compared results for a group of students on the PISA 2015 CPS assessment framework (these students did not participate in PISA 2015), using human and computer-based agents as collaboration partners. Their results "indicated no significant differences between the type of collaboration partner" (p. 1) demonstrating that computer-based agents can proxy human ones. In conclusion, CPS is able to measure an important part of students' collaboration skills and the results provide valuable information to policy makers.

[34]The current, when writing this, Corona Virus Crisis has in an even more acute way showcased the need for proficiency in electronic communication and collaboration skills.

Table 3.4 Collaborative Problem Solving skills evaluated in PISA 2015.

		Collaborative problem-solving competencies		
		(**1**) Establishing and maintaining a shared understanding	(**2**) Taking appropriate action to solve the problem	(**3**) Establishing and maintaining team organisation
Problem-solving processes	(**A**) Exploring and understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve the problem
	(**B**) Representing and formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describing roles and team organisation (communication protocol/rules of engagement)
	(**C**) Planning and executing	(C1) Communicating with team members about the actions to be/being performed	(C2) Enacting plans	(C3) Following rules of engagement (e.g. prompting other team members to perform their tasks)
	(**D**) Monitoring and reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

Notes: The table is based on Figure V.2.1 in OECD (2017b, p. 50).

4. Empirical strategy

The ideal way to estimate the effects of teaching practices on CPS would be to conduct an experiment. In such an experiment, students and teachers would be randomly assigned to different kinds of teaching practices (compare Project STAR that randomly assigned students to small or normal sized classes, see Krueger (1999) for details). Random assignment of teaching practices would ensure unbiased estimates, unaffected by self-selecting students and teachers. In reality, however, students (or their parents) do not choose schools randomly. If students that have higher general intelligence, or have higher Openness and Agreeableness, self-sort into schools with a higher emphasis on modern teaching practices, the estimated effect of modern teaching practices on CPS, without any controls, would be biased upwards. Similarly, if generally better or more motivated teachers seek out schools that emphasize more on traditional teaching practices, the effect of modern teaching practices would typically be biased downwards. This would be the case if teacher motivation, or other variables, affects how well students perform in CPS through other channels than teaching practices. Teachers may also adjust their teaching based on the ability or composition of their students. Teachers with unruly and/or lower ability students might opt to increase the use of modern teaching practices, again leading to a downwards bias in the estimated effect on CPS. Another source of bias would be if there are differences in the implementation of modern teaching practices between schools that employ the same amount of modern teaching. Correctly implemented modern teaching practices might be positive, but "an empirical analysis that is based on the actual average implementation of this teaching practice might not reveal any positive effects." (p. 374 Schwerdt and Wuppermann 2011).

Furthermore, our information on teaching practices, which is based on in-class time use reported by teachers, does not allow us to distinguish between different implementations of the teaching practices. One worry might be that especially the implementation of problem-based teaching differs substantially within our sample. Differences between the ideal and the actual implementation of interactive teaching styles that involve more problembased teaching might also reconcile our findings with supportive evidence of modern approaches to teaching (Lou et al., 1996; Machin & McNally, 2008). Thus, while a certain teaching practice may be very effective if implemented in the correct way, an empirical analysis that is based on the actual average implementation of this teaching practice might not reveal any positive effects. Our results, therefore, do not call for more lecture style teaching in general. (p. 374 Schwerdt and Wuppermann 2011)

Many of these variables that might lead to self-sorting are observable and covered in the PISA context questionnaires, which means that they can be included in the estimations and therefore be controlled for.[35] Some are, however, unobservable and the exclusion of these from the estimations might introduce bias if they are correlated to what teaching practices are employed. The most common approach to try and overcome these sources of bias in previous research has been to exploit within student, between-subject variation and include student fixed effects in the empirical model (Algan et al. 2013; Bietenbeck 2014; Rivkin and Schiman 2015). Student fixed effects require that students' test scores and teacher practices are measured either at different points in time or in different subjects. Since the PISA data is in nature cross-sectional, and teaching practices and CPS are only measured once, it does not allow for estimation strategies such as student fixed effects that would eliminate part of the possible omitted variables bias.

Despite the increasingly common use of student fixed effects to isolate causality in similar research, many researchers still employ OLS plus student, school, and, when available, teacher controls (e.g. Caro et al. 2016; Fuchs and Wößmann 2008; O'Dwyer et al. 2015). In some cases they do so because of the unavailability of data that permits the use of student fixed effects, and sometimes despite this strategy being available. Following this strain of research due to data limitations, a large and rich set of control variables at the student, teacher, and school levels (see Table A11 in Appendix VI), as well as country fixed effects, were included in the OLS estimations. The hope is that this approach will alleviate and minimize the potential omitted variable biases. Nevertheless, the results presented in Section 5 should be interpreted with caution and the estimates are not claimed to be causal. Such estimates are ultimately correlations conditional on the covariates, with some probable bias induced by remaining unobserved variables. Still, they are an important part of a broader picture and give meaningful information about the relationship between teaching practices and CPS.

While estimates based on OLS plus controls might not be able to make causal claims in most cases, they do not suffer from the problems associated with student fixed effects estimations (discussed in chapter 5.1 in Angrist and Pischke 2009):

1. Increased bias caused by measurement errors since the variation is restricted to within individuals.[36] This typically leads to attenuation bias.

2. Since the estimated coefficients come only from individuals that change treatment status, which might be very few, the internal validity is restricted to this subsample.

Angrist and Pischke (2009) therefore advise that "it's important to avoid overly strong

[35] Their inclusion should also help shrink the gap between results based on teacher-reported and student-reported teaching practices (Desimone et al. 2010).

[36] For this reason, OECD recommends against using student fixed effects when analyzing the PISA data, see https://www.oecd.org/pisa/data/httpoecdorgpisadatabase-instructions.htm

claims when interpreting fixed-effects estimates" (p. 169).[37] OLS estimations with controls might therefore not be such a bad idea, as long as the limitations are acknowledged; especially when analyzing data sets such as PISA or TIMSS.

Based on the discussion above, the estimation equation of the education production function explaining CPS ability is estimated with OLS and has the following form:

$$
\begin{aligned}
CPS_{isc} = \alpha + \beta_1 TradTI_{sc} + \beta_2 ModTI_{sc} + \beta_5 Sch_{sc} + \beta_3 Stu_{isc} + \beta_4 Tch_{sc} \\
+ C_c + \beta_6 D^{stu}_{isc} + \beta_7 D^{tch}_{sc} + \beta_8 D^{sch}_{sc} + \varepsilon_{isc}
\end{aligned}
\tag{4.1}
$$

where CPS_{isc} is the CPS test score of student i, in school s and country c. It is determined by the school average traditional teaching practice index ($TradTI_{sc}$) and the school average modern teaching practice index ($ModTI_{sc}$), as well as by vectors of control variables at the school (Sch_{sc}), student (Stu_{isc}), and teacher (Tch_{sc}) levels. C_c is the country fixed effects and ε_{isc} is the student-specific error term. D^{sch}_{sc}, D^{stu}_{isc}, and D^{tch}_{sc} are vectors with dummy variables indicating missing values for the school, student, and teacher level control variables. For each student, each of these dummy variables is equal to 0 if the corresponding control variable was not missing and is equal to 1 if it was missing.

The parameters of interest are β_1 and β_2. First, the model in Equation (4.1) was estimated with only either $TradTI_{sc}$ or $ModTI_{sc}$ included, gradually adding country fixed effects, and school, student, and teacher controls as well as the corresponding missing value dummy variables. In the next step, the same model was estimated, but this time with both teaching practice indices included, adding controls in the same way. This procedure was performed for the indices constructed from both the student and the teacher questionnaires.[38]

Thus, coefficients where CPS was regressed on only one index measures the effect on CPS of an increase in the percentage of lessons in which traditional (modern) teaching practices are used, letting modern (traditional) practices vary. When both indices were included as explanatory variables, the coefficients reflect the effect on CPS of an increase in the percentage of lessons in which traditional (modern) teaching practices are used, holding modern (traditional) teaching practices constant. In this case, teaching practices not deemed traditional nor modern can be decreased to make room in the lesson for either more traditional or modern activities. Another possibility is that the length of the lessons

[37] Furthermore, Desimone et al. (2010) argues that claims regarding causality should not be made based on non-experimental data. There is also the problem with researchers using student fixed effects, claiming causal relationships, while not taking into account the survey design with plausible values and sampling weights in studies such as TIMSS and PISA (for an overview of this problem, see Jerrim et al. 2017).

[38] Considering external validity from the sample of 17 countries, Equation (4.1), for the student-reported teaching practice indices, was also estimated with the sample including all countries that had the CPS option. This sample consisted of 40 countries and 391,888 students.

could be increased.

Note that the specification in Equation (4.1) estimates the effect of traditional and modern teaching practices to be the same for all students and schools but with different country-specific intercepts. In this thesis the level of analysis is at the individual student, pooling all students in all countries into one sample. Other approaches of conducting analysis are at the country level (Bishop 1997) or the student level within each country (Caro et al. 2016; OECD 2017b), limiting variation to either between countries or within countries. It might be interesting to explore the full heterogeneity of the effect of the teaching practice indices on CPS by performing the analysis within each country. However, when pooling all students into one sample and including country fixed effects, the degrees of freedom of the analysis increases vastly. In this way, a common estimate for all countries, without country-specific differences, is obtained. Nonetheless, heterogeneity between groups of countries is explored in Section 5.4.

A noteworthy aspect of the PISA data in regards to the specification in Equation (4.1) is that PISA does not sample classrooms as TIMSS does. The consequence is that teachers at a school can not be connected to specific students within that school. This means that it is not possible to disentangle within-school variation in teaching practices between classrooms from the school level aggregate (see discussion in Caro et al. 2016 and O'Dwyer et al. 2015). Thus, the measure becomes blunter and might mask differences within a school. Consider a school that has two teachers where one employs more modern teaching practices, and another employs more traditional. The school aggregate will show an average of the two and the regression estimates will not be able to connect students' different CPS scores to the different teachers' teaching practices. This is, however, more of a precision problem, and with enough variation between schools (Echazarra et al. 2016), sufficient precision in the estimates will be ensured. Furthermore, the school average, as compared to the classroom average, of teaching practices is less likely to be endogenous after relevant covariates have been controlled for (Lavy 2016, p. 94).

Concerning the estimation of Equation (4.1), the sampling method used by PISA, with the primary sampling unit being schools, implies that independence at the level of the individual student could not be assumed. In the regressions reported in Section 5, standard errors were therefore clustered at the school level, assuming independence at this level (Fuchs and Wößmann 2008; Avvisati and Keslair 2014). Furthermore, the design of PISA, with student weights and plausible values for the standardized CPS test score, described in Section 3, was accounted for by using the OECD developed Repest package for Stata (Avvisati and Keslair 2014).[39] Specifically, all reported coefficients are averages

[39]Note that the repest command in Stata automatically clusters standard errors at the school level (Avvisati and Keslair 2014). For more detailed information on the analytical framework for PISA 2015, see OECD (2017a) and the PISA homepage at http://www.pisa.oecd.org.

from 10 estimations using each of the 10 CPS plausible values together with the final student weights; 800 more estimations from all combinations of the 80 student weight replicates and the 10 plausible values were used to compute standard errors (for details, see chapter 7 and 8 in OECD 2009).[40]

4.1 Further discussion of biases

Since a pooled sample of students from 17 countries was used in the regressions, country-specific factors that are correlated with teaching practices might introduce bias. A country might have a culture, or long-standing institutions, that promote communication, cooperation, and collaboration, independent of teaching practices. Then, if modern teaching practices are employed more frequently in this country, compared to other countries, the estimated effect of modern teaching practices on CPS will be biased upwards. One such example, highlighted as a significant challenge by O'Dwyer et al. (2015), is the significant employment of supplementary instruction outside of school hours in countries like Japan and South Korea. The inclusion of country fixed effects takes care of these issues, ensuring no bias induced by time-invariant country factors.[41] Moreover, in contrast to O'Dwyer et al. (2015) a measure of total instruction time in all subjects as well as a measure of outside of school instruction time is included as controls (see Table A11 in Appendix VI), ensuring that differences within countries are accounted for.

Regarding time variability in institutional factors, they are unlikely to have a large effect and can be assumed to be highly time persistent. It could have been a problem since the measured teaching practices will reflect how teachers have taught during at least the last year, and possibly for the last 3 years. If institutions changed more rapidly this might introduce bias. Institutional changes are however generally slow and gradual, implying that such factors are "rather similar, during a student's life in secondary school (Fuchs and Wößmann 2008, p. 222). Bias from this factor is therefore unlikely to be a significant problem.

An often overlooked problem with education production function estimates is whether the impact of the explanatory variable of interest on the dependent variable varies over time (Fuchs and Wößmann 2008, p. 222). Even though internal validity might be high, external validity in time will be low if the effect varies. An example would be that, due to some technological invention that enables students to learn much faster and more efficiently, the effect of more instruction time on test scores falls drastically.[42] For teaching

[40]For more detailed information on the methodology and how to analyze the PISA database, see OECD (2009) and https://www.oecd.org/pisa/data/httpoecdorgpisadatabase-instructions.htm.

[41]Variation in teaching practices will, however, be restricted to between schools, net of the average country differences.

[42]A famous and important such relationship in economics that has changed over time is the Philips

practices, their effect on CPS might change based on how much emphasis is placed on teaching them during a teachers' education. Furthermore, the quality of how they are taught might also have an effect. It is likely that if countries place more focus on imprinting teachers with the ability to teach communication and cooperation, or if CPS skills become more permeated in society, the effect of using different teaching practices on CPS scores will change. It is therefore important to continuously re-examine relationships to see if they have changed. The present differences in these factors between countries and regions are, however, mostly controlled for with the inclusion of country fixed effects and the battery of control variables.

Another possibility to control for students' unobserved ability, apart from student fixed effects and regardless of the cross-sectional design of PISA, would be if student's prior achievement could be included as a control variable. These data are, however, seldom available in large-scale international student assessment tests (O'Dwyer et al. 2015), and neither in PISA 2015. Another strategy would be to include student's mathematics or science PISA test scores as controls (recall that the correlation coefficient between test score in CPS and mathematics was 0.75 and 0.82 between CPS and science). For indices based on student-reports this approach would not be advisable since teaching practices have been found to independently affect these test scores (Bietenbeck 2014; Goldhaber and Brewer 1997; Lavy 2016), making them dependent variables and thus bad controls (see chapter 3.2.3 in Angrist and Pischke 2009).[43] However, teacher-reported teaching practices have seldom been shown to have a substantial, statistically significant effect on student test scores in mathematics and science (Algan et al. 2013; O'Dwyer et al. 2015; Schwerdt and Wuppermann 2011). These findings are corroborated by the results from regressions based on Equation (4.1), estimating the relationship between teaching practices and test scores in mathematics and science (see Table A3–A5 in Appendix III). Nevertheless, in Section 5.5 the robustness of the results to the inclusion of the mathematics or science test score as a control in the regressions is investigated.

5. Results

Table 5.1 and 5.2 report different OLS estimations of the empirical specification in Equation (4.1). The dependent variable in all regressions was the standardized CPS test score, and the main explanatory variable, teaching practices, was either the traditional or the modern index, or both. In Table 5.1 and 5.2, the results are also divided by whether the teaching practices were reported by students (Panel A) or by teachers (Panel B).

curve.

[43] See also Cinelli et al. (2020) for further discussions on bad, neutral, and good controls.

5.1 Student-reported teaching practice indices

The results based on the student-reported indices indicate a strong and statistically significant relationship between traditional teaching practices and CPS test scores. The coefficients are all statistically significant at the 0.1 % level, but they decrease in strength with the inclusion of control variables (though it increases slightly when only country dummies are added) and lands at 1.2 when country dummies, as well as school, student, and teacher controls, are included (column 5 in Table 5.1). This pattern suggests an upward bias, based on observable variables, in the effect of traditional teaching practices on CPS. When the traditional index is the only treatment variable, see columns 1–3 in Panel A of Table 5.2, the estimated coefficient is significantly lower in the cases when no, only country dummies, or both country dummies and school controls were included. Though, this disparity decreases to a difference of only 0.1 when student and teacher controls are included in columns 4 and 5.

Table 5.1 CPS and teaching practices, both traditional and modern indices at the same time.

	(1)	(2)	(3)	(4)	(5)
Panel A: Student-reported indices					
Traditional teaching index	3.441***	3.536***	2.509***	1.316***	1.201***
	(0.157)	(0.188)	(0.161)	(0.156)	(0.165)
Modern teaching index	-2.318***	-2.235***	-1.618***	-0.667***	-0.693***
	(0.144)	(0.198)	(0.189)	(0.179)	(0.188)
Panel B: Teacher-reported indices					
Traditional teaching index	-0.003	0.489	0.125	-0.070	-0.150
	(0.176)	(0.272)	(0.234)	(0.177)	(0.170)
Modern teaching index	0.342	0.213	0.193	0.107	0.075
	(0.189)	(0.219)	(0.163)	(0.121)	(0.123)
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes
Observations	119,702	119,702	119,702	119,702	119,702

Notes: All regressions include a constant. The dependent variable in all regressions is the standardized plausible values for CPS. Control variables are listed in Table A11 in Appendix VI. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weight replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5.2 CPS and teaching practices, traditional and modern indices included separately.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Student-reported indices										
Traditional teaching index	2.361***	2.773***	1.850***	1.213***	1.097***					
	(0.153)	(0.184)	(0.160)	(0.154)	(0.160)					
Modern teaching index						-0.744***	-1.040***	-0.700***	-0.498**	-0.545**
						(0.152)	(0.210)	(0.188)	(0.174)	(0.183)
Panel B: Teacher-reported indices										
Traditional teaching index	0.160	0.601**	0.223	-0.018	-0.119					
	(0.148)	(0.214)	(0.204)	(0.156)	(0.155)					
Modern teaching index						0.340*	0.417*	0.244	0.078	0.019
						(0.159)	(0.165)	(0.139)	(0.105)	(0.110)
Country dummies	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes	No	No	No	No	Yes
Observations	119,702	119,702	119,702	119,702	119,702	119,702	119,702	119,702	119,702	119,702

Notes: All regressions include a constant. The dependent variable in all regressions is the standardized plausible values for CPS. Control variables are listed in Table A11 in Appendix VI. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weight replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In contrast, in Table 5.1 the student-reported modern teaching practice index seems to almost mirror the behavior of the traditional index with a minus sign. The coefficients start highly negative at -2.3 when no controls are included. As more controls are added in columns 2–4, the coefficients decrease in absolute value and end up at -0.69 when all controls are included in column 5. The pattern thus suggests a downward bias in the effect of modern teaching practices on CPS test score. When the traditional teaching practice index is not included, and the modern index is the only main explanatory variable, a similar pattern is observed. Columns 6-10 in Table 5.2 report these results. Similar to when the traditional index is the only treatment variable included, coefficients are generally lower (in absolute terms) than when both indices are included, especially in columns 6–8 with none or few added controls. The pattern and significance levels, however, are mostly the same: all coefficients are statistically significant at the 0.1 % or 1 % level, and although they do not decrease in absolute value in every step as more controls are added, the coefficient is less negative in columns 9 and 10 compared to columns 6–8. Notably, when all controls are included the coefficient is very close to the one when both indices are included. Accordingly, the correlation between teaching practices and CPS test score in the full specification of Equation (4.1), conditional on the covariates, seems rather stable even though the other index can vary in the specifications where only one index is included.

An interesting observation is that coefficients change substantially when both school controls and student controls are added. Teacher controls, by contrast, seem to only have a marginal effect on point estimates for the indices,[44] though a low effect of teacher controls is in line with similar previous research (e.g. Bietenbeck 2014; Blazar 2015). Additionally, adding country dummy variables in column 2 of Table 5.1 has almost no effect on the point estimates, indicating low heterogeneity in the relationship between countries. When the indices are included separately, however, in Table 5.2 there is a larger change in the coefficients.

The most comprehensive and complete specification is the one in column 5 of Table 5.1, where both traditional and modern indices, as well as country dummies and controls at the school, student, and teacher levels, are included. This is also the specification where the most meaningful interpretations based on the results can be made. The coefficient for the traditional index implies that an increase of one standard deviation in the index (ca. 11 %), holding the modern index constant, is related to a 12.8 % of a standard deviation increase in the CPS test score.[45] In contrast, the coefficient for the modern teaching practice index indicates that an increase of one standard deviation in this index (ca. 11 %), holding the traditional index constant, is associated with a 7.4 % of a standard deviation decrease in CPS test score.

The 21^{st}-century skills movement recommends more modern teaching practices to foster CPS skills in students. However, the results presented thus far indicates the opposite; that modern teaching practices are related to lower CPS test scores, and that traditional practices are related to higher scores. Critics of the 21^{st}-century skills movement and modern teaching practices would argue that students' general academic ability is fostered by the inculcation of facts-based knowledge and that CPS skills are developed in tandem with cognitive skills (Kay 2010; Rotherham and Willingham 2010). Focusing less on instilling facts-based knowledge and more on modern teaching activities could harm this development, thus explaining the observed results. Although these results are not claimed to be causal estimates, they indicate that moving towards more modern, and away from traditional, teaching might harm student's CPS abilities. However, it should also be remembered that results based on student-reported teaching practices should be

[44]It is worth noting that the teacher controls are school averages and therefore essentially school controls. School controls are thus made up of the principal-reported and the teacher-reported variables.

[45]Recall that the standardized CPS test score has a standard deviation of 1, implying that if the traditional teaching practice index would go from 0 to 1, the corresponding associated increase in the CPS test score would be 1.2 standard deviations. Translating this into standard deviations for the teaching practice index too results in the reported figure.

interpreted more carefully than those based on teacher-reports.

5.2 Teacher-reported teaching practice indices

Panel B in Table 5.1 and 5.2 report the results from regressions using the teacher-reported teaching practice indices. Compared to the student-reported indices discussed above, a completely different pattern emerges. In all specifications where both indices are included in the regression, the coefficient for both indices is small and they are never statistically significant.[46] When only one index is included in Table 5.2, the coefficients are comparable to when both indices are included in terms of both magnitude and statistical significance. The exceptions are when the modern index attains significance at the 5 % level in columns 6 and 7, and when the traditional index attains significance at the 1 % level when only country dummies are included in column 2.[47] Notably, despite being statistically insignificant, the coefficients for the modern and traditional indices are flipped, in columns 4 and 5 of Table 5.1, compared to when the student-reported teaching practices were used; the modern index has a positive coefficient and the traditional has a negative. The same flipping of signs occurs in Table 5.2.

In the most comprehensive and complete specification in column 5 of Table 5.1, the traditional teaching practice index has a coefficient of -0.150 and the corresponding coefficient for the modern index of 0.075. These coefficients imply that an increase of one standard deviation in the traditional (modern) index, holding the modern (traditional) index constant, is associated with a decrease (increase) of 0.017 (0.008) standard deviations in CPS test score.[48] Both effects are indicated to be very small and the coefficients are statistically insignificant. Thus, in contrast to the results based on student-reports, using teacher-reported teaching practices indicates a zero effect on CPS test scores.

Lastly, comparing the results based on student-reported teaching practices in Table A1 and A2 in Appendix III and Table 5.1 (Panel A), it appears that the results in the reduced sample, with only the 17 countries that took the teacher questionnaire option, is a good representation of the whole sample of 51 countries that took the CPS test. Coefficients are comparable, except for the index for modern teaching practices when not using any controls, with the expected smaller standard errors when using the whole sample. The external validity, in terms of extrapolating the results to all countries that had the CPS option, is therefore also likely to be high for the results based on the teacher-reported

[46] Though when only country dummies are included the coefficient for the traditional index increases from -0.003, to 0.489.

[47] This implies a probably larger heterogeneity between countries for the traditional index compared to the modern one, which is partially explored in Section 5.4.

[48] Standard deviations for the traditional and modern indices were both ca. 11 %.

teaching practice indices, reported in Panel B of Table 5.1 and 5.2.[49]

5.3 Discussion

The striking difference in results, based on whether student- or teacher-reported teaching practices are used, roughly corresponds to previous findings in the literature of relating teaching practices to student test scores in mathematics and science. First, let us consider the results based on student-reported teaching practices, reported in Panel A of Table 5.1 and 5.2. The statistically significant and positive association between the traditional index and CPS is similar to, but larger than, those found by Bietenbeck (2014) and Lavy (2016) for science and mathematics test scores. In contrast, Echazarra et al. (2016) identified a zero effect of teacher-directed instruction. Regarding the negative relationship between modern teaching practices and CPS, they corroborate the findings in Goldhaber and Brewer (1997) and Echazarra et al. (2016), although Bietenbeck (2014) found a zero effect[50] and Lavy (2016) observed a small positive effect. OECD (2017b) examined the association between "communication intensive" teaching practices[51] and CPS test scores. They found a statistically significant negative relationship that, like in the results presented here, decreases in absolute value when more controls are added.

Similarly, the observed zero effect of both teacher-reported teaching practice indices, reported in Panel B of Table 5.1 and 5.2, are also essentially corroborating those in the previous literature when relating science or mathematics test score to teaching practices. Algan et al. (2013) found that teaching practices were unrelated to students' cognitive performance, and O'Dwyer et al. (2015) concluded that both traditional and modern teaching practices explained very little of the variability in mathematics test scores. Schwerdt and Wuppermann (2011) also identified a zero effect for modern teaching practices, though a statistically significant positive effect of traditional teaching practices was observed, albeit quite small. Recall that Algan et al. (2013), O'Dwyer et al. (2015), and Bietenbeck (2014) used the same data set, TIMSS 2007, though they did not consider precisely the same teaching practice variables, and they did not employ the same empirical method.[52] The differences in their results, based on whether student- or teacher-reported teaching

[49]It could, however, be the case that students from the 51 countries are more uniform across countries than teachers are, making the sample of 17 countries not as representative for teachers.

[50]This was the case for the overall test score; when the reasoning part of the test was the dependent variable, the effect was slightly positive.

[51]Two of the four questions included in this teaching practice are included in the modern teaching practice index in this thesis: "Students are given opportunities to explain their ideas" and "There is a class debate about investigations".

[52]O'Dwyer et al. (2015) used multilevel regression models with controls, Algan et al. (2013) used normal OLS with controls, but included school fixed effects, while Bietenbeck (2014) employed a student fixed effects approach.

practices were used, are nonetheless noteworthy, and they correspond to the findings in this thesis in regards to CPS.

The results regarding the relationship between teaching practices and student test scores in mathematics and science (see Table A3–A5 in Appendix III) also followed the same patterns, corroborating the findings in the previous literature.[53] Specifically, for the student-reported indices, when all controls where included, the traditional index had a large and statistically significant coefficient, while the modern index had a statistically significant negative coefficient. For the teacher-reported teaching practices, statistically significant results were observed with no or few control variables. However, when all control variables were included no statistically significant relationships remained, and coefficients were close to zero.

Interestingly, for both student- and teacher-reported teaching practices, the coefficients were comparable in size to the ones in the regressions with CPS as the dependent variable. This raises the question of how well CPS manages to isolate collaboration skills from general problem solving skills. Despite CPS supposedly being more of a non-cognitive skill, the results here indicate that they are determined mostly in the same way as general cognitive skills (recall also the high correlation between CPS test scores and mathematics and science test scores, see Section 3.4).

It is recognized that the empirical method employed is presumably not able to control for all omitted variable biases. Results are ultimately correlations conditional on the covariates and should be interpreted with caution. No strong policy recommendations are therefore made based on the results, although they are hopefully able to act as a stepping stone for future research on the subject. Bearing this in mind, the inclusion of such a rich set of control variables at the student, school, and teacher levels likely produced results fairly close to the true relationships. The relatively high degree of correspondence between the results in this thesis regarding CPS, mathematics, and science test scores and those in the literature relating teaching practices to proficiency in mathematics and science, indicates this further. Moreover, the fact that results did not change even when students mathematics/science test score was included as a control (proxying general ability), in the regressions with teacher-reported indices, gives additional credence to their reliability. See Section 5.5 for a discussion on this robustness test.

In summary, the reported relationships between teaching practices and student cognitive skills in the literature seem to systematically differ based on whether student- or teacher-reported teaching practices are used in the analysis. After reviewing the available

[53] The regressions where only one of the teaching practice indices were included at a time, approximately followed the same patterns observed when CPS was the dependent variable. With no or few controls, compared to when both indices were included, coefficients were significantly smaller in absolute value. However, when all controls were included, the point estimates were very close and significances were roughly the same.

literature on the subject, Desimone et al. (2010) concluded that teacher-based answers are less biased, less prone to measurement errors, and therefore more reliable as proxies for the "true" teaching practices employed. This interpretation implies that there is no relationship between what teaching practices teachers use and students' CPS test scores.

The difference in results could, however, be the effect of the student- and teacher-reported indices measuring different things based on how the teaching practice questions were formulated (recall that the questions that constituted the modern teaching practice index were not precisely the same in both questionnaires, see Table 3.2). This could also explain the discrepancy in results between Algan et al. (2013) and O'Dwyer et al. (2015) on one hand, and Bietenbeck (2014) on the other. Nonetheless, the fact that the questions comprising the traditional teaching practice index were identical in both questionnaires, while still generating significantly different coefficients in the regressions, indicates that it is less likely to be the case that the results are driven by how the questions were formulated. The wording in the modern teaching practice questions is also very similar in both questionnaires and should, therefore, measure approximately the same thing. Moreover, despite the identical traditional teaching practice questions, the correlation between the teacher-reported index and the student-reported index were as low as 0.43 (which is in line with previous research, see Section 2.3). This constitutes further evidence against the interpretation that different things were measured in the student- and teacher-reported indices because of how the questions were formulated. Instead, the implication is that teachers and students have different *perceptions* of what teaching practices are used.

The statistically significant relationships observed in regressions based on student-reported indices, indicate that student perception of what teaching practices teachers employ is related to their CPS test scores. What could explain this systematic relationship? As mentioned in Section 2.3, if teaching practices reported by students simply have higher random measurement errors, the consequence should be attenuation bias; which is the opposite of what the results indicate. Therefore, there has to be some systematicity in the student-reported teaching practices that could explain the results. One possible explanation could be that students to some degree grade their teachers when answering the questions on teaching practices. If this is the case, better teachers are generally perceived by students as more traditional in their teaching, and worse teachers are considered as employing more modern practices. The inclusion of such a rich set of teacher control variables should, however, mitigate this tendency. Nevertheless, the determinants of these student perceptions must be examined further to try and understand how they are determined and how they can be influenced. An ideal approach would be a smaller, more qualitative study, over a longer period, involving several ways to try and discern how these perceptions are formed and influenced. Such a study could for example employ deep interviews with both students and teachers at certain time intervals, and independent observations of teachers teaching their students. The results of such studies could then help inform how to better proceed when constructing and analyzing large student assessments such as PISA. Regardless, it would seem unreasonable to base education policy solely on student perceptions, especially when they differ so much from teacher-reports.

In conclusion, the most reasonable interpretation of the results is that the true teaching practices employed by teachers are unlikely to be related to CPS test scores.[54] At the very least, it is highly implausible that modern teaching practices have a positive effect on CPS and that traditional teaching practices have a negative effect. The conclusion that modern teaching practices have either a zero effect or a negative effect on CPS test scores would also discredit the calls made by the 21^{st}-century skills movement for more modern teaching practices as a way of developing collaboration skills in students. The indication that traditional teaching practices have either a zero effect or a positive effect on CPS further discredits the 21^{st}-century skills movement's policy recommendations on this subject. Conversely, modern teaching practices might have other positive effects, documented in research based on teacher-reported teaching practices. They have for instance been connected to the development of social capital in students (Algan et al. 2013).[55] It is, however, not advisable to increase modern teaching practices based solely on this evidence, especially if traditional teaching practices are decreased at the same time. On the contrary, although no strong recommendations are made, results in this thesis indicate that if policy makers want to increase students' CPS skills they should, if anything, increase traditional teaching practices and decrease modern teaching practices in schools.

5.4 Heterogeneity

To gain further insight into the relationship between teaching practices and CPS, heterogeneity of the results was examined across several dimensions commonly investigated in the literature. In this section and the next, the heterogeneity and robustness of the regressions in Table 5.1 are discussed, though approximately the same patterns were observed for the specifications where the indices were used one at a time.[56] For all of the examined dimensions, the results based on teacher-reported teaching practices were essentially homogeneous, with a few exceptions that are mentioned below[57].

First, the sample was divided across different country groups; both OECD compared

[54]Students' CPS skills might, like their cognitive skills, instead be influenced by factors like class size (Angrist and Lavy 1999; Krueger 1999) and instruction time (Cattaneo et al. 2017). Future research is therefore urged to examine these relationships.

[55]If the results based on teacher-reported teaching practices are more correct, then modern teaching practices affect neither cognitive skills (Algan et al. 2013; O'Dwyer et al. 2015) nor CPS skills, and expanding their use might not be such a bad idea. Policy makers are, nonetheless, advised to be cautious and this indication is something future research will have to try and ascertain.

[56]All these regressions are available upon request.

[57]The regressions where heterogeneity was observed are included in Appendix IV.

to non-OECD countries and Western countries compared to Asian and South American countries.[58] The results did not differ substantially between OECD and non-OECD countries, though some differences were observed with no or only country controls. However, when all controls were included, qualitatively and quantitatively the same results were observed.

When differentiating between Western, Asian, and South American countries, heterogeneity was observed for the student-reported indices, see Table A6 in Appendix IV. Modern teaching practices were less negative in South America and the Western countries (also less precisely estimated, which could be the result of the lower sample sizes) and more negative in the Asian countries, compared to the results in Table 5.1. Thus, the negative association in the full sample predominantly comes from the Asian countries, indicating that the student perception of increasing modern teaching practices in these education systems might be more harmful to student CPS abilities.[59]

Next, the sample was divided between boys and girls but, interestingly, no differences between the groups were observed. This implies a homogeneous relationship between teaching practices and CPS abilities based on sex, which is unexpected since previous research has found differences in how boys and girls learn (e.g. Nasser 2016). It is, however, in line with what others have found when relating instructional hours to science or mathematics test scores (Lavy 2015). Further, as mentioned in Section 3.4, PISA intentionally reduced the difficulty to low or intermediate on the individual problem solving abilities needed on the CPS tests. Thus the CPS tests should primarily measure collaborative, and not cognitive, skills. Nonetheless, there is a risk that the relationship between teaching practices and CPS is different in the lower and upper part of the cognitive ability distribution. This concern turned out to be unfounded since no significant differences were observed between the coefficients when dividing the sample between low-achieving and high-achieving students, first in CPS and then in science.[60]

The last three dimensions where heterogeneity was examined were the type of school (public vs private), natives vs immigrants, and socio-economic status (SES). No notable differences were observed for the teacher-reported teaching practices on these dimensions, except for a larger coefficient, though still statistically insignificant, for the traditional in-

[58] The OECD countries were Australia, The Czech Republic, Germany, Spain, Italy, Portugal, The United States of America, Chile, and South Korea. The non-OECD countries were Hong Kong, China-Taipei, Beijing-Shanghai-Jiangsu-Guangdong, Macau, Brazil, Colombia, Peru, and the United Arab Emirates. The Western countries were Australia, Czech Republic, Germany, Spain, Italy, Portugal, and The United States of America. Asian countries were Hong Kong, South Korea, China-Taipei, Beijing-Shanghai-Jiangsu-Guangdong, and Macau. Lastly, South American countries were Brazil, Chile, Colombia, and Peru.

[59] Future research is encouraged to further investigate the heterogeneity of the effect of teaching practices on CPS between different countries and country groups.

[60] The division was performed by setting the threshold between low- and high-achieving at the average test score in each subject.

dex for second-generation immigrants. Results based on student-reported teaching practices differed more substantially. See Table A7 and Table A8 in Appendix IV for the regression results on school type and natives vs immigrants. With all controls included, the absolute value of the coefficients for both the traditional and the modern index was larger for the private independent schools, compared to the government-dependent private schools and public schools. Consequently, the relationship seems to be more negative for the student perception of more modern teaching practices in independent private schools. For both first- and second-generation immigrant students, the significance of both the traditional and modern index disappeared when student and teacher controls were added[61]. Further, the coefficient for the traditional index was significantly lower, compared to the 1.2 for native students, at 0.1 for the first-generation and 0.5 for second-generation immigrants. Thus, the teaching practices used (as perceived by students) does not seem to be related to immigrant students' CPS test scores, though the coefficients for the second-generation immigrants are closer to those of the native students.[62] Lastly, when comparing low SES students with high SES students[63], the coefficient for the traditional index was slightly larger and the modern index was slightly more negative in most specifications[64].

In summary, the relationship between teaching practices and CPS seems highly homogeneous, especially when teacher-reported teaching practices were used. The exceptions were some differences between country groups, between school forms and between levels of SES, as well as major differences between native and immigrant students.

5.5 Robustness of results

A viable concern is that the reported results are the effect of the specific decisions made throughout this thesis when it comes to things such as sample restrictions, setting missing values to zero, and the definition of teaching practices. Therefore, to test the sensitivity of the results, numerous robustness tests were performed based on these decisions as well as the decisions made on which control variables to include and what grades students included in the analysis were in.[65]

[61]Though the loss of significance could be the result of a much lower sample size for both first-generation (8,248) and second-generation (8,718) immigrant students.

[62]This conclusion is contrary to the results in Lavy (2015) when relating instructional time to science or mathematics test scores. He found larger positive effects of more instructional time for first and second-generation immigrants, compared to natives.

[63]The division was performed in the same way as for CPS and Science test scores.

[64]Specifically when all controls were included the traditional index for the high SES students had a coefficient of 1.372 (compared to 0.985 for the low SES students) and the modern index was -0.750 (compared to -0.631).

[65]These regressions are available upon request, though some are included in Appendix V (mostly those where results were not indicated to be robust).

The first sensitivity test was performed by not making the sample restrictions described in Appendix I so that the whole sample of 142,235 students in 4,707 schools were used.[66] Results were robust to not making the sample restrictions; coefficients typically differed only slightly and all significances except one were the same.[67]

Further, the approach of setting all missing values for the control variables to zero could bias the results if values are not missing at random. To test this concern re-estimations of the regressions were performed, this time without the missing value dummies for the control variables (compare Fuchs and Wößmann 2008). This specification implicitly assumes that observations are missing conditionally at random, and should, therefore, produce the same results as in the full model based on Equation (4.1). All results were quantitatively and qualitatively virtually the same[68], confirming no major systematicity in the missing values.

As discussed in Section 4.1 a possible way to control for students unobserved ability is to use students' mathematics or science test scores as a control variable. However, these variables are likely to be bad controls and would in this context be expected to induce attenuation bias, were they to be included. This is especially the case in the regressions were student-reported teaching practices were used since teaching practices based on student-reports have been found to independently affect test scores in mathematics and science (Bietenbeck 2014; Goldhaber and Brewer 1997; Lavy 2016). Estimations in this thesis indicate the same relationships. However, for the teacher-reported teaching practices, there were no statistically significant relationships when other factors were controlled for. Nonetheless, as a robustness check, the main specification in Table 5.1 was re-estimated with the test score in mathematics/science as a control variable.

When teacher-reported indices were used, the results did not change qualitatively compared to the main model in column 5 of Table 5.1, though coefficients were even less significant with much larger standard errors. However, when the student-reported indices were used, the inclusion of science test score as a control removed the significances and reduced the point estimates absolute values significantly. The coefficient for the traditional index was reduced to 0.053 and the modern index to -0.153. With mathematics test score included as control, the significances stayed at the 1 % level for the traditional index and

[66] Note that this is the sample where students are still restricted to only those in schools that had both students and teachers that answered the questions on teaching practices (if this restriction is not imposed, the sample using student-reported indices would be 153,846 students in 5,354 schools). The reason for this restriction is, like before, to ensure comparability of the results based on whether they were based on teaching practices reported by students or by teachers.

[67] The exception was the teacher-reported traditional index with only country dummies, that was significant at the 5 % level, which could be explained by the larger sample size enabling the model to be able to better pinpoint the relationship

[68] The results for the complete model including both teaching practice indices are reported in Table A9 in Appendix V

the 5 % level for the modern index. The corresponding point estimates were also larger in absolute value; 0.433 for the traditional index and -0.348 for the modern index.

The question is whether these results are trustworthy or just a manifestation of attenuation bias induced by a bad control. It is difficult to ascertain and the best option is for them to be compared to ones made using a method that more confidently can claim to control for unobserved student ability. Nevertheless, if the outcomes are not the result of bad control variables, it effectively makes the results based on teacher- and student-reported teaching practices practically the same: no statistically significant relationship between teaching practices and CPS. Thus, the relationship to CPS for traditional teaching practices is either positive or zero, and either negative or zero for modern teaching practices, with this robustness control giving more credence to the zero effect.

Another possible concern is how teaching practices were defined, see Section 3.3. This is not the only possible way to define teaching practices and, therefore, other definitions were explored (compare Bietenbeck 2014). See Table A10 in Appendix V for the regression results. First, instead of creating the modern and traditional indices by assigning values between zero and one to the four possible answers to the teaching practice questions, the indices were based on the share of answers in a school equal to "Many lessons" or "Every lesson or almost every lesson". The indices were thus created by first calculating, for each question, the percentage of students/teachers in a school that answered "Many lessons" or "Every lesson or almost every lesson".[69] These percentages were then averaged across all modern (traditional) questions to create the modern (traditional) index.

The regressions in which these alternative indices were used as main explanatory variables produced the same results qualitatively and in regards to significance, though coefficients were lower in absolute value for most specifications. Generally, the results based on the teacher-reported teaching practices were closer to the corresponding results reported in Section 5 than those based on the student-reported ones. Notably, when both teacher-reported indices were included all significances except one were the same (insignificant).[70]

A second possible way to define teaching practices is to use the gap between the traditional and modern index as the treatment variable. The gap is a measure of the relative emphasis a school puts on traditional teaching in relation to modern teaching. If the gap is positive the school employs traditional teaching practices in a higher percentage of lessons, while if it is negative a higher percentage of modern teaching practices are used. Results based on both student-reported and teacher-reported teaching practices,

[69] This was true for all questions except the modern index questions in the student questionnaire, where the corresponding answers were "In most lessons" and "In all lessons".

[70] The exception was for the traditional index when only country dummies were included. In this case, the coefficient was significant at the 0.1 % level, indicating a larger heterogeneity between countries than previously observed.

when the gap was used as the main explanatory variable, were qualitatively the same as those reported in Section 5[71].

Finally, following Fuchs and Wößmann (2008), sensitivity to grade inclusion and certain control variables were explored. The main regressions were re-estimated without some variables that could be argued to be outcomes of what teaching practices were used, i.e. bad controls, and arguably superfluous variables for parents' education[72]. The results were virtually the same when teacher-reported teaching practices were used. For the student-reported indices, results were qualitatively the same, but the absolute values of the coefficients for both indices were somewhat larger, compared to the main specifications in Section 5. If these variables are bad controls, they induce attenuation bias, albeit quite small. Lastly, there is a concern that teaching practices might affect 15-year-old's differently based on what grade the student attended, especially for grades far above or below the national modal grade. Alternative specifications were therefore estimated in which the sample was restricted to only those 15-year-old's attending the national modal grade for 15-year-old students, plus-minus one grade. However, results were robust to this restriction and coefficients were practically the same as in the main specifications.

Overall the results were remarkably robust to most restrictions and decisions made in this thesis. The major exception was when student-reported teaching practices were used, and students' science/mathematics test score was included as a control. Although this changes the interpretation of the results, the principal conclusion does not change: traditional teaching practices have a positive or zero effect on CPS while modern teaching practices have a negative or zero effect.

[71]For the student-reported teaching practices, coefficients were large and statistically significant at the 0.1 % level while coefficients were close to zero and statistically insignificant for the teacher-reported teaching practices.

[72]The student level variables excluded were Enjoyment of science, Science self-efficacy, Instrumental motivation, Achieving motivation, Mother's education level, Father's education level, Disciplinary climate in science classes, Teacher support in science classes, Perceived feedback, and Adaption of instruction. Also, two teacher level controls were removed: Job satisfaction and Satisfaction with the teaching profession.

6. Conclusion

There is an ongoing debate on on what teaching practices should be used to foster the skills needed for the current and future workplace. This thesis has presented empirical evidence that documents the connection between teaching practices and students' collaboration ability. Using the PISA 2015 data on teaching practices and CPS test scores, OLS regressions with country dummies as well as controls at the student, teacher, and school level were used to investigate this relationship. Results differed significantly based on whether the student- or teacher-reported teaching practices were used in the regressions. When student-reported indices were used, traditional teaching practices were positively related to CPS test scores, with large coefficient estimates. More modern practices, on the contrary, were observed to have a negative association with CPS test scores.

In contrast, when teacher-reported teaching practices were included in the regressions, the relationship to CPS was zero. Since teacher-reported teaching practices are plausibly more reliable than student-reported ones (Desimone et al. 2010) teaching practices are unlikely to affect CPS test scores. Or, if they do have an effect, as indicated by the regressions using student-reported teaching practices, modern teaching practices is indicated to negatively affect collaboration ability. In conclusion, the relationship between traditional teaching and CPS is either zero or positive, while it is either zero or negative for modern teaching. The call for more modern, and less traditional, teaching practices to foster 21^{st}-century skills, such as collaboration skills, is therefore discredited.

Interestingly, from the results based on student-reported teaching practices, student *perception* of what practices are employed were indicated to affect CPS. Future research should, therefore, investigate the determinants of student perceptions of what teaching practices are used and why they differ from the teacher reported ones. Smaller, more qualitative approaches are recommended in this regard. The discrepancy in results, in this thesis and previous research, based on who answered the questionnaire should also be examined further. Moreover, despite these discrepancies, future research is urged, when possible, to use both student- and teacher-reported measures to get a richer, more comprehensive and complete picture. Investigating other factors that might affect CPS skills, such as class size, instruction time, and the use of Information and communication technology (ICT), is also encouraged. Further, it is desirable, when possible, to use a larger set of countries for the teacher-reported teaching practices to examine the external validity of the results from the 17 countries in this thesis.

Although the data used in this thesis did not allow for student fixed effects estimations, so that possible endogeneity in the estimates could have been reduced, the results are nonetheless likely to be quite reliable. The inclusion of a large battery of controls, several robustness tests, and the similarities with results in previous research, indicate that the presented results are probably close to the true relationships. The possible remaining endogeneity is still a concern that prevents any identified relationships to be interpreted as causal. Future research should, therefore, when it is possible, try to overcome this problem. Hopefully, data on teaching practices and CPS that allows for student fixed effects, or other strategies to control for unobserved variables, will become available for future research. Nevertheless, the evidence presented illuminates part of a broader picture and brings some clarity to what constitutes good teaching, and what determines student CPS ability. Finally, acknowledging the limitations, policy makers are still urged to be cautious in disregarding traditional teaching practices and be wary of implementing more modern teaching practices in schools.

References

ACT, Inc (2013). *Student- and Teacher-Reported Behavioral Measures: Do They Agree?* Issue Brief. URL: https://eric.ed.gov/?id=ED546856.

Algan, Yann, Pierre Cahuc, and Andrei Shleifer (July 2013). "Teaching Practices and Social Capital". In: *American Economic Journal: Applied Economics* 5.3, pp. 189–210. URL: http://www.aeaweb.org/articles?id=10.1257/app.5.3.189.

Ananiadou, Katerina and Magdalean Claro (2009). *21st Century Skills and Competences for New Millennium Learners in OECD Countries*. OECD Education Working Papers No. 41. OECD Publishing, Paris. URL: https://www.oecd-ilibrary.org/content/paper/218525261154.

Anderson, Elizabeth A. (2019). "K-12 Online Student Engagement: New Construct or Extension to Extisting Construct?" In: *Handbook of Research on Emerging Practices and Methods for K-12 Online and Blended Learning*. Ed. by T.L. Heafner, R. Hartshorne, and R. Thripp. Advances in Early Childhood and K-12 Education (2329–5929). IGI Global, pp. 155–171. URL: https://books.google.se/books?id=3ayMDwAAQBAJ.

Angrist, Joshua D. and Victor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". In: *The Quarterly Journal of Economics* 114.2, pp. 533–575. URL: http://www.jstor.org/stable/2587016.

Angrist, Joshua and Jorn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. 1st ed. Princeton University Press. URL: https://EconPapers.repec.org/RePEc:pup:pbooks:8769.

Avvisati, Francesco and François Keslair (Sept. 2014). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values*. Statistical Software Components, Boston College Department of Economics. URL: https://ideas.repec.org/c/boc/bocode/s457918.html.

Bietenbeck, Jan (2014). "Teaching practices and cognitive skills". In: *Labour Economics* 30, pp. 143–153. URL: http://www.sciencedirect.com/science/article/pii/S0927537114000219.

Bill and Melinda Gates Foundation (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. MET project, Research Paper. Seattle, WA, USA: Bill and Melinda Gates Foundation. URL: http://k12education.gatesfoundation.org/resource/learning-about-teaching-initial-findings-from-the-measures-of-effective-teaching-project-2/.

Bishop, John H. (1997). "The Effect of National Standards and Curriculum-Based Exams on Achievement". In: *The American Economic Review* 87.2, pp. 260–264. URL: http://www.jstor.org/stable/2950928.

Blazar, David (2015). "Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement". In: *Economics of Education Review* 48, pp. 16–29. URL: http://www.sciencedirect.com/science/article/pii/S0272775715000710.

Buckley, Katie and Sara Bartolino Krachman (2016). *Patterns in Student Self-Report and Teacher Report Measures of Social-Emotional Mindsets, Skills, and Habits, Initial findings from the Boston Charter Research Collaborative*. Working Paper. Transforming Education. URL: https://www.transformingeducation.org/wp-content/uploads/2017/04/TE-BCRCWorkingPaperFINAL.pdf.

Caro, Daniel H., Jenny Lenkeit, and Leonidas Kyriakides (2016). "Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012". In: *Studies in Educational Evaluation* 49, pp. 30–41. URL: http://www.sciencedirect.com/science/article/pii/S0191491X15300286.

Cattaneo, Maria A., Chantal Oggenfuss, and Stefan C. Wolter (2017). "The more, the better? The impact of instructional time on student performance". In: *Education Economics* 25.5, pp. 433–445. eprint: https://doi.org/10.1080/09645292.2017.1315055. URL: https://doi.org/10.1080/09645292.2017.1315055.

Cinelli, Carlos, Andrew Forney, and Judea Pearl (Mar. 2020). *A Crash Course in Good and Bad Controls*. URL: https://www.researchgate.net/publication/340082755_A_Crash_Course_in_Good_and_Bad_Controls.

Common Core Standards (2020). URL: http://www.corestandards.org/read-the-standards/ (visited on 06/11/2020).

Core Knowledge foundation (2020). URL: https://www.coreknowledge.org/ (visited on 06/11/2020).

Desimone, Laura, Thomas Smith, and David Frisvold (Mar. 2010). "Survey Measures of Classroom Instruction, Comparing Student and Teacher Reports". In: *Educational*

Policy - EDUC POLICY 24, pp. 267–329. URL: https://journals.sagepub.com/doi/10.1177/0895904808330173.

Donné, Noémie Le, Pablo Fraser, and Guillaume Bousquet (2016). *Teaching Strategies for Instructional Quality: Insights from the TALIS-PISA Link Data*. OECD Education Working Papers No. 148. OECD Publishing, Paris. URL: https://www.oecd-ilibrary.org/content/paper/5jln1hlsr0lr-en.

Echazarra, Alfonso et al. (2016). *How teachers teach and students learn: Successful strategies for school*. OECD Education Working Papers No. 130. OECD Publishing, Paris. URL: https://www.oecd-ilibrary.org/content/paper/5jm29kpt0xxx-en.

Fuchs, Thomas and Ludger Wößmann (2008). "What accounts for international differences in student performance? A re–examination using PISA data". In: ed. by Machin S. Dustmann C. Fitzenberger B. The Economics of Education and Training. Studies in Empirical Economics. Physica-Verlag HD, pp. 209–240.

Goldhaber, Dan D. and Dominic J. Brewer (1997). "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity". In: *Journal of Human Resources* 32.3, pp. 505–523. URL: https://EconPapers.repec.org/RePEc:uwp:jhriss:v:32:y:1997:i:3:p:505-523.

Herborn, Katharina et al. (2020). "The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans?" In: *Computers in Human Behavior* 104, p. 105624. URL: http://www.sciencedirect.com/science/article/pii/S0747563218303571.

Hughes, Richard L. and Steven K. Jones (2011). "Developing and assessing college student teamwork skills". In: *New Directions for Institutional Research* 2011.149, pp. 53–64. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ir.380. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ir.380.

Jerrim, John et al. (2017). "What happens when econometrics and psychometrics collide? An example using the PISA data". In: *Economics of Education Review* 61, pp. 51–58. URL: http://www.sciencedirect.com/science/article/pii/S0272775717300869.

Kay, Ken (2010). "21st Century Skills: Why They Matter, What The are, and How We Get There". In: *21st Century Skills: Rethinking How Students Learn*. Ed. by James A. Bellanca and Ronald S. Brandt. Bloomington: Solution Tree Press.

Krueger, Alan B. (May 1999). "Experimental Estimates of Education Production Functions". In: *The Quarterly Journal of Economics* 114.2, pp. 497–532. eprint: https://academic.oup.com/qje/article-pdf/114/2/497/5218380/114-2-497.pdf. URL: https://doi.org/10.1162/003355399556052.

Lau, Kwok-chi and Terence Yuk-ping Lam (2017). "Instructional practices and science performance of 10 top-performing regions in PISA 2015". In: *International Journal of Science Education* 39.15, pp. 2128–2149. eprint: https://doi.org/10.1080/09500693.2017.1387947. URL: https://doi.org/10.1080/09500693.2017.1387947.

Lavy, Victor (2015). "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries". In: *The Economic Journal* 125.588, F397–F424. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12233. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12233.

— (Mar. 2016). "What Makes an Effective Teacher? Quasi-Experimental Evidence". In: *CESifo Economic Studies* 62.1, pp. 88–125. eprint: https://academic.oup.com/cesifo/article-pdf/62/1/88/6998337/ifv001.pdf. URL: https://doi.org/10.1093/cesifo/ifv001.

Lucas, Bill (May 2019). *Why we need to stop talking about twenty-first century skills.* URL: https://www.researchgate.net/publication/332864663_Why_we_need_to_stop_talking_about_twenty-first_century_skills.

Nasser, Amany (Dec. 2016). *The Difference Between Girls and Boys in Learning.* URL: https://www.researchgate.net/publication/311855897_The_Difference_Between_Girls_and_Boys_in_Learning.

NCTM (1991). *Professional standards for teaching mathematics. Report.* Reston, VA: National Council of Teachers of Mathematics. URL: https://scholar.google.com/scholar_lookup?title=Professional%5C%20standards%5C%20for%5C%20teaching%5C%20mathematics%5C&author=NCTM%5C&publication_year=1991.

Nelson, Laurie Miller (1999). "Collaborative Problem Solving". In: *Instructional-design theories and models. A new paradigm of instructional theory. Volume II.* Ed. by Charles M. Reigeluth. Mahwah, N.J.: Lawrence Erlbaum Associates, pp. 241–266.

NRC (1996). *National science education standards. Report.* Washington, DC: National Research Council. URL: https://scholar.google.com/scholar_lookup?title=National%5

C%20science%5C%20education%5C%20standards%5C&author=NRC%5C&publication_year=1996.

O'Dwyer, Laura M., Yang Wang, and Katherine A. Shields (2015). "Teaching for conceptual understanding: A cross-national comparison of the relationship between teachers' instructional practices and student achievement in mathematics". In: *Large-scale Assessments in Education* 3.1. URL: https://link.springer.com/article/10.1186/s40536-014-0011-6#citeas.

OECD (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. Paris: OECD Publishing. URL: https://www.oecd-ilibrary.org/content/publication/9789264056275-en.

— (2013). "The Skills Needed for the 21st Century". In: *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. OECD Publishing, Paris. URL: https://www.oecd-ilibrary.org/content/component/9789264204256-5-en.

— (2017a). *PISA 2015 Assessment and Analytical Framework*. Paris: OECD Publishing. URL: https://www.oecd-ilibrary.org/content/publication/9789264281820-en.

— (2017b). *PISA 2015 Results (Volume V)*. Paris: OECD Publishing. URL: https://www.oecd-ilibrary.org/content/publication/9789264285521-en.

— (2017c). *PISA 2015 Technical Report*. Paris: OECD Publishing. URL: https://www.oecd.org/pisa/data/2015-technical-report/.

Partnership for 21st Century Learning (2020). *P21 Network*. URL: https://www.battelleforkids.org/networks/p21 (visited on 06/08/2020).

Partnership for 21st Century Skills (2020). *Partnership for 21st Century Skills (P21)*. URL: https://www.imls.gov/assets/1/AssetManager/Bishop%5C%20Pre-Con%5C%202.pdf (visited on 08/07/2020).

Rivkin, Steven G. and Jeffrey C. Schiman (2015). "Instruction Time, Classroom Quality, and Academic Achievement". In: *The Economic Journal* 125.588, F425–F448. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12315. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12315.

Rotherham, Andrew J. and Daniel T. Willingham (2010). ""21st Century" Skills: Not New, but a Worthy Challenge". In: *American Educator* 34.1, pp. 17–20. URL: https://eric.ed.gov/?id=EJ889143.

Schwerdt, Guido and Amelie C. Wuppermann (2011). "Is traditional teaching really all that bad? A within-student between–subject approach". In: *Economics of Education Review* 30.2, pp. 365–379. URL: http://www.sciencedirect.com/science/article/pii/S0272775710001640.

Stadler, Matthias et al. (July 2019). "Computer–Based Collaborative Problem Solving in PISA 2015 and the Role of Personality". In: *Journal of Intelligence* 7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6789457/.

Urdan, Tim (2010). "The challenges and Promise of Research on classroom Goal Structures". In: *Handbook of Research on Schools, Schooling and Human Development*. Ed. by J.L. Meece and J.S. Eccles. Taylor & Francis, pp. 92–109. URL: https://books.google.se/books?id=nt-NAgAAQBAJ.

Wang, Lijuan et al. (Jan. 2009). "Assessing Teamwork and Collaboration in High School Students: A Multimethod Approach". In: *Canadian Journal of School Psychology* 24. URL: https://eric.ed.gov/?id=EJ842790.

Wentzel, Kathryn R. and Katherine Muenks (2016). "Peer Influence on Students' Motivation, Academic Achievement, and Social Behaviour". In: *Handbook of Social Influences in School Contexts: Social-Emotional, Motivation, and Cognitive Outcomes*. Ed. by K.R. Wentzel and G.B. Ramani. Educational Psychology Handbook. Taylor & Francis. URL: https://books.google.se/books?id=sqheCwAAQBAJ.

Wößmann, Ludger and Martin West (2006). "Class–size effects in school systems around the world: Evidence from between–grade variation in TIMSS". In: *European Economic Review* 50.3, pp. 695–736. URL: http://www.sciencedirect.com/science/article/pii/S0014292105000152.

Wu, Margaret (2005). "The role of plausible values in large-scale surveys". In: *Studies in Educational Evaluation* 31.2. Measurement, Evaluation, and Statistical Analysis, pp. 114–128. URL: http://www.sciencedirect.com/science/article/pii/S0191491X05000209.

Zemelman, Steven, Harvey Daniels, and Arthur A. Hyde (2005). *Best practice: Today's standards for teaching and learning in America's schools*. 3rd ed. Heinemann. URL: https://scholar.google.com/scholar?hl=sv&as_sdt=0%5C%2C5%5C&q=Best++practice%5C%3A++Today%5C%%5C%80%5C%99s+standards+for+teaching+and+learning+in+America%5C%E2%5C%80%5C%99s+schools.+3rd%5C&btnG=.

Appendix

I Data cleaning procedures

When creating the student-reported teaching practice indices from the student questionnaire data set, 129 schools (1,530 students) did not have any student that answered the questions making up one or both of the traditional and modern indices and were therefore removed. From this reduced sample, schools with less than 10 students were excluded to guarantee a minimum of precision in the estimate of teaching practices; in this way, the teaching practices in a school were not based on the answers of only a few students (compare Bietenbeck 2014). In this step, 1,205 students were removed from 159 schools. Furthermore, for the same reason, schools with less than 5 students answering the questions on teaching practices were removed (3,378 students from 332 schools).

The same procedures were performed for the teacher questionnaire data set when the teacher-reported teaching practice indices were constructed. 217 schools (3,424 teachers) were removed because these schools did not have any teacher that answered the questions making up one or both of the traditional and modern indices. Additionally, schools with less than 6 teachers, as well as schools where less than 3 teachers answered the questions on teaching practices, were removed. In these steps, 11,991 teachers from 810 schools and 33 teachers from 8 schools were removed, respectively. Then, all teacher control variables, as well as the traditional and modern teaching practice indices, were averaged in each school. Since all teacher control variables were school averages, schools with less than 3 teachers answering these questions were removed (923 teachers from 59 schools).

Next, the student questionnaire data set was merged with the teacher questionnaire data set and the data set containing students' test scores in mathematics, science, and CPS, using the school identifier provided by PISA. When applying the exclusion restrictions in the student and teacher questionnaire data sets, the same schools were not always removed. Consequently, to ensure that the students were the same in the regressions using both student-based and teacher-based teaching practices, the final sample was reduced further. Specifically, 29,561 students from 1,149 schools were excluded in those schools where there where not any, or too few, teachers answering the teaching practice questions.

II Xandar

In Figure A1 an example of the CPS computer interface that students faced is shown. In this particular item, the student has to interact with the computer agents Alice and Zach to solve problems regarding the fictional country Xandar.

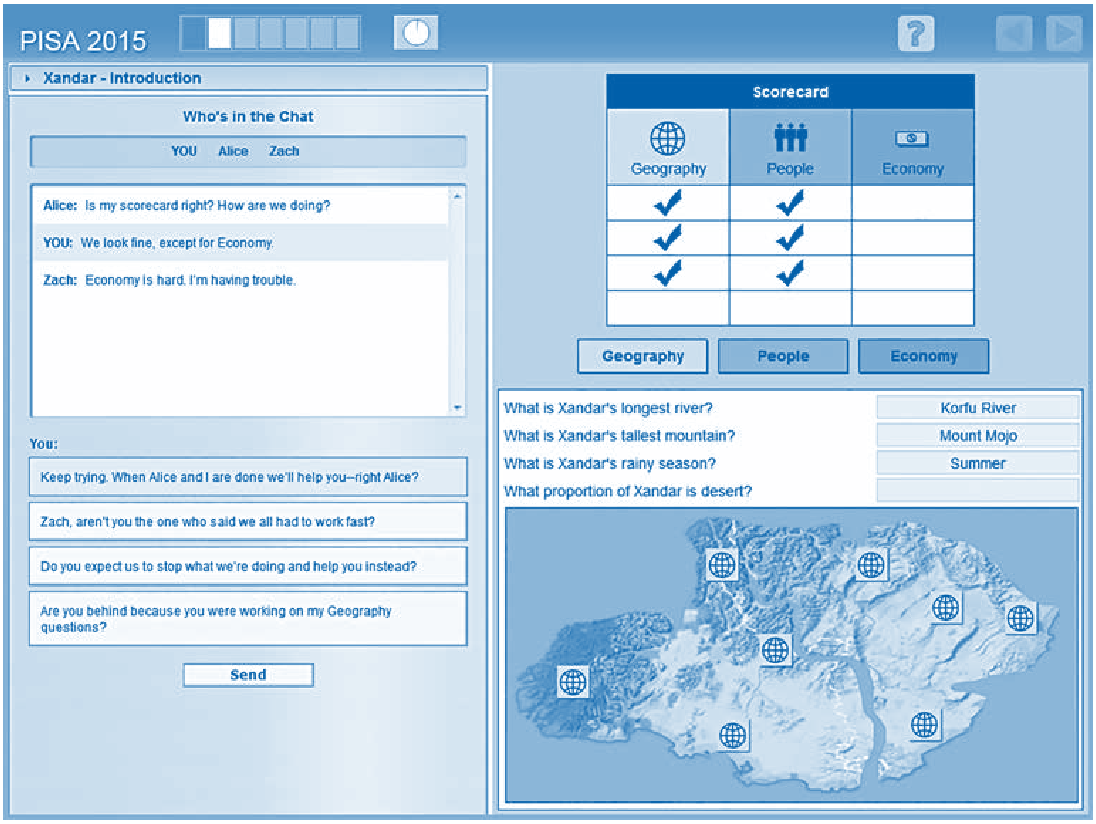

Figure A1 Screenshot of "Figure V.2.17, XANDAR: Part 4, Item 2" from OECD (2017b, p. 62).

III Additional regression results

Table A1 CPS and teaching practices for all countries with the CPS option, using student-reported indices.

	(1)	(2)	(3)	(4)
Traditional teaching index	3.212***	3.142***	2.267***	1.200***
	(0.093)	(0.102)	(0.097)	(0.090)
Modern teaching index	-3.147***	-2.477***	-1.750***	-0.701***
	(0.079)	(0.104)	((0.099)	(0.089)
Country dummies	No	Yes	Yes	Yes
School controls	No	No	Yes	Yes
Student controls	No	No	No	Yes
Observations	391,888	391,888	391,888	391,888

Notes: All regressions include a constant. The dependent variable in all regressions is the standardized plausible values for CPS. Control variables are listed in Table A11 in Appendix VI. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A2 CPS and teaching practices for all countries with the teacher questionnaire option, using student-reported indices.

	(1)	(2)	(3)	(4)	(5)
Traditional teaching index	3.629***	3.407***	2.394***	1.275***	1.179***
	(0.139)	(0.156)	(0.148)	(0.128)	(0.129)
Modern teaching index	-2.509***	-2.311***	-1.636***	-0.652***	-0.678***
	(0.121)	(0.155)	(0.155)	(0.134)	(0.141)
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes
Observations	149,263	149,263	149,263	149,263	149,263

Notes: All regressions include a constant. The dependent variable in all regressions is the standardized plausible values for CPS. Control variables are listed in Table A11 in Appendix VI. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A3 Mathematics/science test scores and teaching practices for all countries with the CPS option, using student-reported indices.

	(1)	(2)	(3)	(4)
Panel A: Mathematics				
Traditional teaching index	3.156***	3.494***	2.597***	1.313***
	(0.102)	(0.122)	(0.122)	(0.101)
Modern teaching index	-3.558***	-2.326***	-1.623***	-0.533***
	(0.086)	(0.111)	(0.107)	(0.092)
Panel B: Science				
Traditional teaching index	3.524***	3.750***	2.813***	1.473***
	(0.095)	(0.117)	(0.116)	(0.091)
Modern teaching index	-3.443***	-2.585***	-1.857***	-0.706***
	(0.079)	(0.105)	(0.103)	(0.079)
Country dummies	No	Yes	Yes	Yes
School controls	No	No	Yes	Yes
Student controls	No	No	No	Yes
Observations	391,888	391,888	391,888	391,888

Notes: All regressions include a constant. The dependent variable in all regressions is the standardized plausible values for either mathematics or science. Control variables are listed in Table A11 in Appendix VI. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A4 Mathematics/science test scores and teaching practices, using the final sample and student-reported indices.

	(1)	(2)	(3)	(4)	(5)
Panel A: Mathematics					
Traditional teaching index	3.134***	3.835***	2.746***	1.263***	1.122***
	(0.181)	(0.206)	(0.205)	(0.175)	(0.171)
Modern teaching index	-3.357***	-2.039***	-1.410***	-0.444**	-0.503**
	(0.163)	(0.194)	(0.178)	(0.164)	(0.164)
Panel B: Science					
Traditional teaching index	3.690***	4.182***	3.077***	1.591***	1.461***
	(0.151)	(0.197)	(0.189)	(0.138)	(0.140)
Modern teaching index	-2.952***	-2.300***	-1.679***	-0.632***	-0.687***
	(0.145)	(0.189)	(0.183)	(0.158)	(0.160)
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes
Observations	119,702	119,702	119,702	119,702	119,702

Notes: All regressions include a constant. Control variables are listed in Table A11 in Appendix VI. The dependent variable in all regressions is the standardized plausible values for either mathematics or science. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A5 Mathematics/science test scores and teaching practices, using the final sample and teacher-reported indices.

	(1)	(2)	(3)	(4)	(5)
Panel A: Mathematics					
Traditional teaching index	-1.492***	0.454	0.117	-0.122	-0.180
	(0.191)	(0.288)	(0.236)	(0.164)	(0.156)
Modern teaching index	0.457*	0.445	0.388*	0.234	0.179
	(0.210)	(0.241)	(0.165)	(0.128)	(0.134)
Panel B: Science					
Traditional teaching index	-0.773***	0.400	0.065	-0.184	-0.251
	(0.175)	(0.278)	(0.236)	(0.156)	(0.154)
Modern teaching index	0.614**	0.449	0.395*	0.244*	0.174
	(0.189)	(0.230)	(0.159)	(0.107)	(0.105)
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes
Observations	119,702	119,702	119,702	119,702	119,702

Notes: All regressions include a constant. Control variables are listed in Table A11 in Appendix VI. The dependent variable in all regressions is the standardized plausible values for either mathematics or science. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

IV Heterogeneity regressions

Table A6 CPS and teaching practices for Western, Asian, and South American countries, using student-reported indices.

	(1)	(2)	(3)	(4)	(5)
Panel A: Western countries					
Traditional teaching index	2.116***	3.334***	2.194***	0.987***	0.952***
	(0.208)	(0.312)	(0.261)	(0.210)	(0.215)
Modern teaching index	-1.581***	-2.286***	-1.540***	-0.450	-0.533*
	(0.303)	(0.309)	(0.283)	(0.253)	(0.258)
Observations	48,156	48,156	48,156	48,156	48,156
Panel B: Asian countries					
Traditional teaching index	3.156***	4.367***	2.648***	1.098***	1.025***
	(0.368)	(0.459)	(0.316)	(0.274)	(0.253)
Modern teaching index	-2.939***	-2.527***	-2.188***	-1.168***	-1.234***
	(0.442)	(0.431)	(0.346)	(0.301)	(0.272)
Observations	30,473	30,473	30,473	30,473	30,473
Panel C: South American countries					
Traditional teaching index	3.295***	3.177***	2.181***	1.283***	1.159***
	(0.226)	(0.253)	(0.276)	(0.280)	(0.284)
Modern teaching index	-1.365***	-2.155***	-1.314***	-0.452	-0.354
	(0.197)	(0.306)	(0.294)	(0.278)	(0.270)
Observations	29,295	29,295	29,295	29,295	29,295
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes

Notes: All regressions include a constant. Control variables are listed in Table A11 in Appendix VI. The dependent variable in all regressions is the standardized plausible values for CPS. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7 CPS and teaching practices for private independent, private (government-dependent), and public schools, using student-reported indices.

	(1)	(2)	(3)	(4)	(5)
Panel A: Private Independent					
Traditional teaching index	2.131***	2.364***	1.699***	1.590***	1.585***
	(0.376)	(0.393)	(0.345)	(0.418)	(0.395)
Modern teaching index	-2.420***	-2.318***	-1.926***	-1.258**	-1.165**
	(0.357)	(0.547)	(0.526)	(0.446)	(0.413)
Observations	14,635	14,635	14,635	14,635	14,635
Panel B: Private, government-dependent					
Traditional teaching index	1.399**	2.217***	1.775***	0.845**	0.778**
	(0.454)	(0.369)	(0.334)	(0.307)	(0.279)
Modern teaching index	-1.994***	-2.302***	-1.386***	-0.817**	-0.869**
	(0.428)	(0.367)	(0.295)	(0.288)	(0.289)
Observations	17,141	17,141	17,141	17,141	17,141
Panel C: Public					
Traditional teaching index	3.756***	3.481***	2.412***	1.163***	0.981***
	(0.191)	(0.241)	(0.199)	(0.177)	(0.179)
Modern teaching index	-2.336***	-2.193***	-1.651***	-0.781***	-0.763***
	(0.172)	(0.230)	(0.187)	(0.198)	(0.202)
Observations	75,660	75,660	75,660	75,660	75,660
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes

Notes: All regressions include a constant. Control variables are listed in Table A11 in Appendix VI. The dependent variable in all regressions is the standardized plausible values for CPS. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A8 CPS and teaching practices for natives, first-generation immigrants, and second-generation immigrants, using student-reported indices.

	(1)	(2)	(3)	(4)	(5)
Panel A: Natives					
Traditional teaching index	3.571***	3.573***	2.569***	1.435***	1.319***
	(0.153)	(0.179)	(0.151)	(0.153)	(0.162)
Modern teaching index	-2.431***	-2.209***	-1.641***	-0.690***	-0.706***
	(0.144)	(0.190)	(0.180)	(0.174)	(0.183)
Observations	98,802	98,802	98,802	98,802	98,802
Panel B: First-generation immigrants					
Traditional teaching index	2.332***	2.392***	1.723***	0.039	0.110
	(0.452)	(0.575)	(0.457)	(0.498)	(0.460)
Modern teaching index	-1.274*	-1.693**	-1.198*	-0.665	-0.481
	(0.519)	(0.598)	(0.491)	(0.527)	(0.477)
Observations	8,248	8,248	8,248	8,248	8,248
Panel C: Second-generation immigrants					
Traditional teaching index	2.432***	2.903***	1.207*	0.527	0.526
	(0.428)	(0.603)	(0.548)	(0.574)	(0.555)
Modern teaching index	-1.809***	-1.921***	-1.116*	-0.613	-0.838
	(0.454)	(0.502)	(0.443)	(0.483)	(0.502)
Observations	8,718	8,718	8,718	8,718	8,718
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes

Notes: All regressions include a constant. Control variables are listed in Table A11 in Appendix VI. The dependent variable in all regressions is the standardized plausible values for CPS. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

V Robustness regressions

Table A9 CPS and teaching practices, with both traditional and modern indices at the same time. No missing value dummy variables.

	(1)	(2)	(3)	(4)	(5)
Panel A: Student-reported indices					
Traditional teaching index	3.441***	3.536***	2.521***	1.405***	1.288***
	(0.157)	(0.188)	(0.168)	(0.159)	(0.171)
Modern teaching index	-2.318***	-2.235***	-1.643***	-0.678***	-0.718***
	(0.144)	(0.198)	(0.192)	(0.185)	(0.194)
Panel B: Teacher-reported indices					
Traditional teaching index	-0.003	0.489	0.180	-0.014	-0.111
	(0.176)	(0.272)	(0.239)	(0.188)	(0.181)
Modern teaching index	0.342	0.213	0.117	0.057	0.044
	(0.189)	(0.219)	(0.168)	(0.130)	(0.137)
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes
Observations	119,702	119,702	119,702	119,702	119,702

Notes: All regressions include a constant. The dependent variable in all regressions is the standardized plausible values for CPS. Control variables are listed in Table A11 in Appendix VI. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A10 CPS and alternative teaching practice indices, as well as the gap between the traditional and modern index.

	(1)	(2)	(3)	(4)	(5)
Panel A: Alternative indices: Share of answers ≥ "Many lessons"					
Student-reported indices					
Traditional teaching index	2.451***	2.375***	1.691***	0.901***	0.829***
	(0.108)	(0.120)	(0.105)	(0.107)	(0.114)
Modern teaching index	-1.583***	-1.535***	-1.099***	-0.399***	-0.414***
	(0.116)	(0.149)	(0.130)	(0.120)	(0.124)
Teacher-reported indices					
Traditional teaching index	0.054	0.441***	0.180	-0.013	-0.066
	(0.100)	(0.128)	(0.123)	(0.092)	(0.086)
Modern teaching index	0.091	0.061	0.037	0.054	0.050
	(0.114)	(0.125)	(0.106)	(0.077)	(0.078)
Panel B: Gap between traditional and modern index					
Student-reported indices					
Traditional index minus modern index	2.870***	2.880***	2.024***	0.956***	0.917***
	(0.130)	(0.154)	(0.145)	(0.130)	(0.142)
Teacher-reported indices					
Traditional index minus modern index	-0.163	0.063	-0.064	-0.091	-0.109
	(0.162)	(0.218)	(0.165)	(0.126)	(0.122)
Country dummies	No	Yes	Yes	Yes	Yes
School controls	No	No	Yes	Yes	Yes
Student controls	No	No	No	Yes	Yes
Teacher controls	No	No	No	No	Yes
Observations	119,702	119,702	119,702	119,702	119,702

Notes: All regressions include a constant. The dependent variable in all regressions is the standardized plausible values for CPS. Control variables are listed in Table A11 in Appendix VI. Each coefficient is the average from estimates using each of the 10 plausible values and the final student weights. 800 estimations from combinations of the 80 student weights replicates and the 10 plausible values were used to calculate the school clustered standard errors which are reported in parentheses. Significance is denoted by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

VI Descriptive statistics

Table A11 Descriptive statistics for all control variables.

	Mean	SD	Min	Max	% missing
Student level controls					
Student Gender (0 = boy, 1 = girl)	0.504	0.500	0	1	0.00
Grade compared to modal grade in country	-0.100	0.750	-4	3	0.51
Age	15.801	0.292	15.3	16.4	0.00
Mother's Education (ISCED)					
None	0.038	0.192	0	1	3.27
ISCED 1	0.066	0.248	0	1	3.27
ISCED 2	0.173	0.378	0	1	3.27
ISCED 3B, C	0.073	0.259	0	1	3.27
ISCED 3A, ISCED 4	0.276	0.447	0	1	3.27
ISCED 5B	0.124	0.329	0	1	3.27
ISCED 5A, 6	0.251	0.433	0	1	3.27
Father's Education (ISCED)					
None	0.038	0.192	0	1	4.75
ISCED 1	0.069	0.253	0	1	4.75
ISCED 2	0.175	0.380	0	1	4.75
ISCED 3B, C	0.071	0.258	0	1	4.75
ISCED 3A, ISCED 4	0.260	0.439	0	1	4.75
ISCED 5B	0.120	0.325	0	1	4.75
ISCED 5A, 6	0.266	0.442	0	1	4.75
Highest parental education (years)	13.024	3.223	3	18	2.58
ISEI of father	45.037	21.959	11	89	17.14
ISEI of mother	46.082	21.977	11	89	28.56
Immigration status					
Native	0.853	0.354	0	1	3.29
First-Generation	0.071	0.257	0	1	3.29
Second-Generation	0.075	0.264	0	1	3.29
Language spoken at home = test language?	0.104	0.305	0	1	1.74
Teacher Fairness (scale)	9.644	3.743	1	24	10.57
Total learning time in school (minutes per week)	1691.301	428.094	0	3,000	15.79
Out-of-School Study Time per week (hours)	19.807	14.874	0	70	16.66
Students' expected occupational status (SEI)	61.632	16.994	10	89	16.72
Duration in early childhood education and care	2.479	1.201	0	8	20.09
Grade Repetition	0.169	0.375	0	1	2.00
Programme designation					
General	0.878	0.328	0	1	0.08
Preparation for vocational studies	0.033	0.179	0	1	0.08
Preparation for labor market	0.046	0.210	0	1	0.08
Modular	0.043	0.202	0	1	0.08

Family wealth (Index)	-0.301	1.236	-7.18	4.43	1.83
Cultural possessions at home (Index)	-0.007	0.953	-1.84	2.63	3.17
Home educational resources (Index)	-0.187	1.060	-4.41	1.18	2.41
ICT Resources (Index)	-0.279	1.120	-3.56	3.51	2.04
Subjective well-being: Sense of Belonging to School (Index)	-0.068	0.953	-3.15	2.64	3.00
Enjoyment of science (Index)	0.173	1.058	-2.12	2.16	9.07
Disciplinary climate in science class (Index)	0.057	0.402	-1.68	1.74	0.05
Teacher support in science class (Index)	0.140	0.375	-1.38	1.42	0.05
Perceived feedback in science class (Index)	0.147	0.366	-1	2.27	0.05
Adaption of instruction in science class (Index)	0.085	0.321	-1.19	1.73	0.05
Instrumental motivation (Index)	0.290	0.957	-1.93	1.74	10.48
Test Anxiety (Index)	0.268	0.942	-2.5	2.55	2.57
Achieving motivation	0.182	0.952	-3.09	1.85	2.74
Parents emotional support (Index)	-0.088	0.976	-3.08	1.1	2.28
Science self-efficacy (Index)	0.113	1.224	-3.76	3.28	11.26
Index of economic, social and cultural status	-0.271	1.086	-7.26	4.07	2.01
School level controls					
Community in which school is located					
Village, hamlet or rural area (< 3 000 people)	0.042	0.202	0	1	6.76
Small town (3 000 to about 15 000 people)	0.155	0.362	0	1	6.76
Town (15 000 to about 100 000 people)	0.269	0.443	0	1	6.76
City (100 000 to about 1 000 000 people)	0.274	0.446	0	1	6.76
Large city (with over 1 000 000 people)	0.260	0.438	0	1	6.76
School Size	1294.079	1098.542	0	11,071	9.31
Class Size	31.151	8.960	13	53	7.32
Number of available computers per student at modal grade	0.713	0.750	0	9.35	11.12
Responsibility for resources (Index)	0.010	1.100	-.795	2.82	1.06
Responsibility for curriculum (Index)	-0.022	1.018	-1.26	1.48	1.06
School Ownership					
Private Independent	0.136	0.343	0	1	10.25
Private Government-dependent	0.160	0.366	0	1	10.25
Public	0.704	0.456	0	1	10.25
Total number of all teachers at school	86.961	70.179	0	1,327	9.56
% of all teachers fully certified	0.786	0.347	0	1	12.81
% of all teachers ISCED LEVEL 5A Bachelor	0.691	0.353	0	1	19.31
% of all teachers ISCED LEVEL 5A Master	0.258	0.305	0	1	26.21
% of all teachers ISCED LEVEL 6	0.053	0.126	0	1	18.62
Student-Teacher ratio	16.313	9.358	1	100	11.58
Creative extra-curricular activities (Scale)	1.945	1.018	0	3	8.11
Curricular development (Index)	0.238	0.984	-4.83	3	8.60
Instructional leadership (Index)	0.110	0.970	-3.97	2.23	9.70

Professional development (Index)	0.139	1.067	-3.81	1.81	9.78
Teachers participation (Index)	0.048	1.044	-3.86	2.4	10.04
Shortage of educational material (Index)	0.009	1.107	-1.32	3.61	8.14
Shortage of educational staff (Index)	0.139	1.112	-1.68	3.72	8.30
Student behaviour hindering learning (Index)	-0.120	1.305	-2.39	3.89	8.41
Teacher behaviour hindering learning (Index)	0.210	1.154	-2.12	4.26	8.44
How often are students assessed? (Scale)					
Mandatory standardized tests	1.985	0.786	1	5	21.46
Non-mandatory standardized tests	2.026	0.898	1	5	16.90
Teacher-developed tests	3.963	1.058	1	5	16.55
Teachers judgmental ratings	3.837	1.360	1	5	16.99
Students are grouped by ability into different classes?					
For all subjects	0.075	0.264	0	1	8.95
For some subjects	0.346	0.476	0	1	8.95
Not for any subjects	0.578	0.494	0	1	8.95
Estimated % of 15-year-old's whose heritage language is different from test language	17.439	29.896	0	100	35.84
Estimated % of 15-year-old students with special needs	7.417	10.364	0	97	23.14
Estimated % of 15-year-old students from socioeconomically disadvantaged homes	28.427	27.287	0	100	16.91
School open to Parental Involvement (Scale)	4.432	0.848	0	5	9.94
Parental participation in School (Scale)	38.319	20.154	0	100	26.49
Proportion of girls at school	0.492	0.173	0	1	12.49
Teacher level controls					
Are you female or male?	0.596	0.185	0	1	0.00
How old are you?	42.629	5.134	28.1	59.8	0.00
Year(s) working as a teacher in total	16.720	4.628	4	33.7	0.00
What is the highest level of formal education you have completed? (Scale)	3.338	0.392	1.78	5	0.00
Employment Status Contract: Permanent position or fixed-term contracts	0.701	0.293	0	1	0.00
Teacher Employment Time: Full time or part time	0.884	0.163	0	1	0.00
Originally trained teacher?	0.502	0.194	0	1	0.00
Education match with what you teach?	0.554	0.160	0	1	0.00
Participated in professional development in the last 12 months	0.985	0.030	.692	1	0.00
Satisfaction with the current job environment (Index)	-0.041	0.490	-1.59	1.45	0.00
Satisfaction with teaching profession (Index)	-0.064	0.435	-1.66	1.53	0.00
Educational material shortage teachers view (Index)	0.029	0.655	-1.37	2.24	0.00
Staff shortage teachers view (Index)	0.047	0.580	-1.62	2.75	0.00
Science teacher collaboration (Index)	0.124	0.608	-2.24	2.19	0.00
Exchange and co-ordination for teaching (Index)	0.034	0.546	-1.61	2.4	0.00

Notes: These statistics are based on the final sample size of 119,702 students. See OECD (2017c) for details on the creation of all indices and explanations for all questions in the questionnaires.