American Educational Research Journal
August 2016, Vol. 53, No. 4, pp. 1054–1085
DOI: 10.3102/0002831216653180

Revisiting the Relationship Between International Assessment Outcomes and Educational Production: Evidence From a Longitudinal PISA-TIMSS Sample

Martin Carnoy
Stanford University
National Research University Higher School of Economics
Tatiana Khavenson
National Research University Higher School of Economics
Prashant Loyalka
Stanford University
William H. Schmidt
Michigan State University
Andrey Zakharov
National Research University Higher School of Economics

International assessments, such as the Program for International Student Assessment (PISA), are being used to recommend educational policies to improve student achievement. This study shows that the cross-sectional estimates behind such recommendations may be biased. We use a unique data set from one country that applied the PISA mathematics test in 2012 in ninth grade to all students who had taken the Trends in International Mathematics and Science Survey (TIMSS) test in 2011 and collected information on students' teachers in ninth grade. These data allowed us to more precisely estimate the effects of classroom variables on students' PISA performance. Our results suggest that the positive roles of teacher "quality" and "opportunity to learn" in improving student performance are much more modest than claimed in PISA documents.

Keywords: educational policy, international tests, opportunity to learn, teacher effects, value-added analysis

Introduction

Cross-national comparisons of international student assessments, such as the Trends in International Mathematics and Science Survey (TIMSS) and especially the Program for International Student Assessment (PISA), are increasingly being used to recommend specific educational policies to improve student achievement (see e.g., OECD, 2010, 2013c; Fuchs & Woessmann, 2004). These large-scale, cross-sectional data sets have been used to recommend, for example, hiring better (or more effective) teachers, the more efficient and equitable distribution of educational resources, increased investment in early childhood education, greater emphasis on formal mathematics, and greater decentralization of school management (Loveless, 2014; OECD, 2010, 2011, 2013c; Schleicher, 2014; Woessmann, Luedemann, Schuetz, & West, 2009).

The intention of this article is to show that the cross-sectional analyses forming the bases of such recommendations can lead to simplified and misleading relationships between student performance and school inputs and organization. We show this by using a unique data set for one country, Russia, which includes ninth-grade students' PISA mathematics results in 2012, individual students' mathematics performance on the TIMSS a year earlier, in 2011, and detailed information on students' ninth-grade teachers and curriculum. With information on students' earlier math achievement and

---

Martin Carnoy is a professor of education and economics in the Graduate School of Education at Stanford University, 485 Lasuen Mall, Stanford, CA 94305, USA; e-mail: *carnoy@stanford.edu*. He is also visiting professor at the Higher School of Economics. His research focuses on broad issues of educational policy in different social and economic contexts. Much of his work is international and comparative.

Tatiana Khavenson is a research associate in the International Laboratory for Educational Policy Analysis, National Research University Higher School of Economics in Moscow. She researches the role of academic achievement and social class in social mobility and how public policy can influence student achievement across countries.

Prashant Loyalka is an assistant professor at the Graduate School of Education and a center fellow at the Freeman Spogli Institute at Stanford University. His research focuses on inequalities in education and on understanding/improving the quality of education in countries such as China, Russia, and India.

William H. Schmidt is professor of education and statistics at Michigan State University. He is a leading expert on mathematics education and researches the role of curriculum and opportunity to learn in improving student learning. He has made major contributions to designing and analyzing the international TIMSS and PISA tests.

Andrey Zakharov is deputy director of the International Laboratory for Educational Policy Analysis, National Research University Higher School of Economics in Moscow. His research focuses on econometric analyses of the processes of schooling. He is currently conducting research on further waves of the longitudinal data used in this article.

detailed data on each student's teacher in ninth grade, we are able to estimate more accurately the effects of classroom variables on students' PISA performance. We show that these effects are much more modest than those in cross-section based studies.

The issue here is not a lack of empirical evidence in the broader literature that such policy recommendations could improve student achievement. For example, a number of studies *do* show that hiring "effective" teachers (one of the OECD policy recommendations) can positively impact student achievement gains (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2006; Carnoy, Chisholm, & Chilisa, 2012; Nye, Konstantopoulos, & Hughes, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Similarly, studies show that teachers with certain qualifications such as more years of teaching experience (Clotfelter, Ladd, & Vigdor, 2007; Rivkin et al., 2005; Rockoff, 2004), educational background (Clotfelter et al., 2007; Darling-Hammond, 2009; Goldhaber & Brewer, 2000; Harris & Sass, 2011; Kukla-Acevedo, 2009), and higher levels of teacher certification (Boyd et al., 2006; Clotfelter et al., 2007; Harris & Sass, 2009) have positive, albeit relatively small, effects on student achievement.

Neither is the issue that claims made on the basis of cross-section international assessment data should be rejected out of hand. For example, studies have used international assessment data to show that in addition to teacher qualifications, policies that increase the coverage and amount of time spent on subject matter, known as increasing "opportunity to learn" or OTL, are positively correlated with student achievement (OECD, 2013c; Schmidt et al., 2001). Specifically, the OECD's 2012 PISA report features an analysis of how OTL in mathematics is positively correlated with PISA mathematics achievement (OECD, 2013c). This new evidence in PISA for the importance of OTL, following on similar findings based on the TIMSS (Schmidt et al., 2001), could provide insight into the impact curriculum has on student performance.

The main issue with using international assessment data to derive claims about educational reform policies lies elsewhere—in the nature of the data the TIMSS and particularly the PISA collect. The data are beset by two fundamental problems we are able to resolve in our study. First, consistent with the TIMSS and PISA's main objective of providing international comparisons of student achievement benchmarks, the TIMSS and PISA scores reflect the accumulated knowledge of a student at one point in time: the end of fourth/eighth grade in the TIMSS and at 15 years old in the PISA. This accumulated knowledge is the result of previous and current school/classroom-related factors such as teacher qualifications and non-school/classroom inputs, such as students' family background (Coleman et al., 1966; White, 1982). Controlling for just students' family background (as both TIMSS and PISA are able to do) makes it more plausible that remaining achievement differences among students are the result of current school/classroom-related factors. However, students with similar family background may still differ in academic ability and previous schooling and non-school experiences that influence their current academic achievement (Todd & Wolpin, 2003). In turn, students with higher initial ability may self-select into higher resource schools and classrooms. Thus, test data at only one point in time may substantially overestimate school/classroom effects because they attribute all of a student's current achievement to current school/classroom resources and do not account for self-selection by teachers and students into "better" classrooms and schools (Rothstein, 2009). Controlling for students' previous school achievement does not resolve all the issues of identifying school resource effects on students' current performance, but it provides far less biased results than attributing current outcomes to current school inputs (Chetty, Friedman, & Rockoff, 2014).

The second potential problem—for PISA—is that it randomly samples a small number of 15-year-olds from each school in each sample and does not sample intact classrooms. Thus, PISA cannot directly identify students with particular teachers and particular classroom conditions. This effectively prevents any analysis of students' PISA performance along a key dimension of the schooling process—the classroom. Further, given that students and teachers are not linked in the PISA sample, PISA did not apply a teacher questionnaire.

The absence of student/teacher linked data in the PISA has not deterred the OECD from making policy recommendations concerning "better" teacher characteristics and classroom practice, such as OTL. Their conclusions rely on analyses that use information on teacher characteristics averaged at the school level (reported by principals) and classroom practices from individual students not linked to particular teachers. But without direct and detailed information on teachers and classroom practices in intact classrooms, estimated effects and their statistical significance may be biased. Data on teachers and their practices derived from principal and student self-reports (e.g., in the PISA) may have greater measurement error than that derived from teacher reports (e.g., in the TIMSS). Aggregate measures of teacher characteristics and classroom practices at the school level do not have the same meaning as the individual-level variables on which they were constructed (Lee, 2000). Specifically, aggregate measures not only represent individual teachers but also the presence of teaching resources at the school level as a whole. Thus, the conclusions that policymakers and researchers draw from cross-sectional PISA data likely overestimate the effects of improving teacher quality and practices in the classroom.

More unbiased estimates can only be achieved by addressing these two problems. There have been attempts to do so with the TIMSS data, using structural modeling (Schmidt et al., 2001) and cross-subject student fixed effects (Van Klaveren, 2011). A longitudinal study in Germany has also tried to address these problems by following up 9th graders in the PISA 2003 sample with a curriculum-based student test in the 10th grade and by testing teachers on their subject matter teaching knowledge (Baumert et al., 2010).

Schmidt et al. (2001) used data from intact eighth-grade class samples available from the TIMSS 1995 to estimate student math outcomes. The TIMSS 1995 survey also tested seventh graders in the same school, so Schmidt et al. were able to partially confront the problem of not having pre-test score measures by controlling for a different cohort's seventh-grade performance in the same school. However, this method was not as satisfactory as ours in estimating teacher effects on students because it could not identify individual student gains associated with eighth-grade teachers.

Van Klaveren (2011) used Dutch 2003 TIMSS data on the same students taking math and physics with different teachers to estimate the effect of a particular teaching style (the amount of time teachers spend lecturing in front of the class) on eighth-grade student performance. This identification strategy closely approaches causality (resolving problems one and two) but has the disadvantage of restricting the variation used to estimate effects to teachers within the same school. It also assumes that a particular classroom practice or teacher characteristic has the same impact on student performance in both subjects (Dee, 2007).

The Baumert et al. (2010) study conducted a one-year follow-up of a sample of German 9th graders in intact "PISA classrooms" that had taken the 2003 PISA math and reading tests. The follow-up included a math test for students (now in 10th grade) as well as a math test and questionnaire for the students' 10th-grade teachers. The estimates focused on the impact of teacher math subject content knowledge (CK) and pedagogical content knowledge (PCK) on student achievement.

Like our study, Baumert et al.'s (2010) is longitudinal and is able to link students and teachers. Yet it also differs from ours in at least two important ways. It has the advantage of collecting data on teacher mathematics knowledge (see e.g., Hill, Rowan, & Ball, 2005), not available in either the TIMSS or PISA surveys (or ours). However, rather than using PISA scores as an outcome measure, as we do, in the German research, PISA score is a control variable when examining the impacts of teacher characteristics on a German curricular standards-based test. Their study therefore does not provide direct evidence on the factors explaining students' PISA performance and, hence, on the possible biases in policy recommendations from the PISA results.

These three studies have presented estimates that have likely reduced bias, but none has directly focused on the bias in standard results from international assessment data. By contrast, our study considers the degree to which reported estimates of the relationship between students' PISA performance and teacher characteristics and practices, such as OTL, are biased and OECD claims based on those estimates overstated. We also test how teacher characteristics and OTL differentially impact the learning gains of different types of students—students with different levels of family resources and students with different levels of initial levels of TIMSS math achievement.

Our study uses unique data from a national sample of Russian students that took the TIMSS test in the eighth grade in the spring of 2011 and to whom we applied the PISA test one year later in the ninth grade in spring, 2012. The data include mathematics achievement results on the same students at two points in time, one year apart. We were able to link information on teachers to student information in eighth grade from the TIMSS survey and from a teacher questionnaire we applied in ninth grade. Eighty-three percent of the eighth graders in the original TIMSS sample (2011) who took the PISA test in ninth grade a year later (2012) had the same teacher in ninth grade as in eighth grade. Our enumerators responsible for the application of the PISA test and ninth-grade survey also reported that they had found almost all students with their eighth-grade class group in ninth grade, as is typical in Russian schools.

Because of the advantages of our data, we are able, for the first time, to estimate PISA performance controlling for students' performance on a baseline test (TIMSS), reducing the bias related to problem one of using cross-section data, and to relate student outcomes on the PISA test directly to resources students face in the classroom, including teacher characteristics and teaching practices reported on teacher questionnaires, reducing the problem two inherent in the PISA survey.

We test the impact of OTL and teacher characteristics using a standard educational production function approach (Boyd et al., 2006; Clotfelter et al., 2007; Coleman et al., 1966; Hanushek, 1986; Schmidt et al., 2001; Todd & Wolpin, 2003). Specifically, we use value-added and a series of recursive equations to model the relationship between PISA mathematics scores and student-, classroom-, and school-level factors. We focus on the contributions of two important classroom factors on PISA mathematics scores: (a) teacher "quality" and (b) OTL.

The results from our more carefully specified models suggest that OECD policy recommendations regarding the positive role that teacher "quality" and OTL play in improving student performance are not misplaced but should be more modest and narrowly defined than the OECD claims. For example, only one of the several measures we use to proxy teacher quality—math teachers with mathematics degrees from universities rather than pedagogical institutes—has a positive impact on ninth-grade students' PISA mathematics score when we control for their eighth-grade TIMSS test, but that effect is relatively small. Similarly, in our estimates, greater student exposure to formal mathematics—used in OECD reports as a key measure for OTL—also has a much smaller effect on PISA scores than in OECD estimates. We also find the positive effects of both these "higher quality" classroom resources on PISA scores are limited to students with middle and higher initial (TIMSS) math scores, suggesting that, contrary to what the OECD suggests, improving teacher quality and OTL could have little benefit to initially lower scoring students. Our results therefore suggest that improving the quality of teachers and increasing formal mathematics teaching may not be useful strategies for reducing the math gap between initially low and higher scoring students.

The rest of the article proceeds as follows. In Section 2, we describe in detail the TIMSS and PISA samples that form the bases of our data and the different types of data we collected in each sample. In Section 3, we discuss our empirical strategy. This includes a discussion of education production functions, our statistical approach, and how we address challenges in identifying model parameters. Section 4 presents a series of results, beginning with estimates of how teacher characteristics and OTL are related to student socioeconomic background, followed by our value-added estimates of teacher and OTL effects on PISA math performance. We also present estimates of the heterogeneity of these effects across student family background levels and across student initial math performance levels. Section 5 discusses the results and draws conclusions regarding policy recommendations drawn by the OECD from the PISA data.

Data

To achieve more unbiased estimates of the effect of math teachers and OTL on student PISA performance, we exploited the timing of the 2012 PISA test one year after the TIMSS test in 2011. The base data for our study was the TIMSS 2011 sample in Russia. This representative sample consists of 4,893 eighth-grade students in 231 intact classrooms in 210 schools in 50 regions. Enumerators surveyed these same students in ninth grade in spring 2012. The ninth-grade students were asked to take the PISA test, and they and their school director took the PISA survey. The enumerators successfully followed up with 90% of the student sample: 4,399 students in 229 classes in 208 schools.

The loss of 10% of the sample at follow-up could be nonrandom and could bias our results. As such, we examine the sensitivity of our results to sample attrition. In particular, we compare mean baseline characteristics (student characteristics, students' family academic resources [FAR], and TIMSS test scores) across the baseline and endline samples. We find no significant differences (*t* tests) in the means of any of the variables between the two samples (2011 and 2012), reducing the chances that the results in our article are biased due to attrition (Table 1). The sample is thus roughly representative of eighth- and ninth-grade students in schools across Russia.

Enumerators also applied a new teacher questionnaire for students' ninth-grade teachers. The questionnaire asked teachers to report their preservice education, focusing on where they received their mathematics training, years of mathematics teaching experience, and their teacher "category."

Table 1
Variable Means and Standard Errors (*SE*), TIMSS Questionnaires, 2011 and 2012 Samples, PISA Questionnaire, 2012 Sample

	Eighth-Grade TIMSS 2011, Mean (*SE*)	Ninth-Grade TIMSS 2012, Mean (*SE*)	Ninth-Grade PISA 2012, Mean (*SE*)
TIMSS score	538.98 (3.56)	538.8 (3.68)	
PISA score			486.49 (4.01)
Student age[a]	14.75 (.01)	14.75 (.01)	15.76 (.01)
Percentage female	48.84 (.01)	49.18 (.01)	
Books in home: 0–10, %	6.15 (.00)	6.32 (.00)	
Books in home: 11–25, %	27.21 (.01)	27.60 (.01)	
Books in home: 26–100, %	35.55 (.01)	35.64 (.01)	
Books in home: 101–200, %	17.39 (.01)	17.47 (.01)	
Books in home: 200+, %	13.41 (.01)	12.68 (.01)	
Books in home: missing, %	0.29 (.00)	0.28 (.00)	
Mother's education: <HS complete, %	8.80 (.01)	8.41 (.01)	
Mother's education: HS complete, %	13.37 (.01)	13.74 (.01)	
Mother's education: postsecondary %	27.50 (.01)	27.67 (.01)	
Mother's education: university complete, %	34.52 (.01)	34.49 (.01)	
Mother's education: grad school, %	2.07 (.00)	1.84 (.00)	
Mother's education missing %	13.75 (.01)	13.86 (.01)	
Percentage of class with BIH > sample median BIH[b]	30.81 (.00)	30.21 (.00)	
Language at home: always Russian, %	82.88 (.01)	82.77 (.01)	
Language at home: missing, %	0.15 (.00)	0.14 (.00)	
School type: regular secondary school, %			83.03 (.01)
School type: gymnasium, %			10.65 (.01)
School type: lyceum, %			4.99 (.00)
School type: educational center, %			1.33 (.00)
Teacher preservice math degree			13.13 (.01)
Teacher preservice math education degree			65.44 (.01)
Teacher preservice no math education			21.43 (.01)
Years teaching this class[c]			3.57 (.03)
Experience in teaching math, years			22.24 (.18)
Teacher category: highest, %			36.59 (.01)
Teacher category: first, %			40.97 (.01)
Teacher category: second, %			16.43 (.01)
Teacher has no category, %			6.00 (.00)
Teacher workload: classes, hours/week			23.46 (.11)
Teacher workload: out-of-classes, hours/week			2.40 (.05)
Teacher workload: administration, hours/week			2.06 (.14)
Exposure applied math (index)[d]			1.92 (.01)
Exposure word problems (index)[e]			1.80 (.02)
Exposure formal math (index)[f]			2.12 (.01)

Source. Russia PISA-TIMSS Survey, 2011–2012.
Note. TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment; HS = high school; BIH = books in the home.
[a]Student age in eighth grade.
[b]*N* = 4,881 in 2011 and 4,389 in 2012.
[c]*N* = 4,179.
[d]Range = 0–3.
[e]Range = 0–3.
[f]Range = 0–4.

According to national education policies in Russia, teachers are paid according to a seniority scale, but they can also submit to a certification process that qualifies them for higher "categories" and that earns them additional salary. Eighty-three percent of the eighth graders in the original TIMSS sample (2011) who took the PISA test in ninth grade a year later (2012) had the same teacher in ninth grade as in eighth grade. Our enumerators responsible for the application of the PISA test and ninth-grade survey reported that they had found almost all students with their eighth-grade class group in ninth grade, as is typical in Russian schools.

Student achievement was measured in several subjects in both the TIMSS (baseline) and PISA (endline). The TIMSS tests measured performance in math and science subjects such as physics, chemistry, biology, and earth sciences. The PISA tests measured student achievement in math, science, and reading. We focus on mathematics achievement, mainly because mathematics was the main subject tested in the 2012 PISA. PISA also only had OTL questions for mathematics.

The TIMSS-PISA questionnaires and additional questions we posed to principals provided rich information on student characteristics, students' family academic resources, and whether the school students attend is "regular" or selective. For example, students were 14.8 years old in eighth grade and 15.8 years old in ninth grade (Table 1). They frequently reported that they had a large number (>100) of books in their home—31% in the TIMSS questionnaire. About 37% also reported that their mothers had completed university or taken graduate work. The mean books in the home and mother's education estimates may seem high, but they reflect how cheap books were in Communist times and the high level of education in Russia at the end of the 20th century. In terms of defining mothers' levels of education and books in the home (BIH), we use the TIMSS rather than PISA BIH and mother's education categories. We do this for two reasons: The categories—especially mother's education—on the PISA student questionnaire are less clear than on the TIMSS questionnaire, and the answers to the eighth-grade TIMSS questionnaire better control for "initial conditions" in our estimation strategy.

In addition to individual student characteristics, we also estimated a relative measure of family resources of students in the classroom—the proportion of students in each eighth-grade class who reported categories of books in the home greater than the sample median books in the home (26–100)[1]—and obtained data on the selectivity of the school attended by students. These student composition factors measured at the school/classroom level appear to be important influences on individual student achievement (Carnoy et al., 2012). According to responses by principals to our ninth-grade school questionnaire, approximately 80% of the students in our sample attended "regular" middle/secondary schools, while about 20% attended elite, selective secondary schools—almost all public and only differing in

Greek name—called gymnasiums and lyceums. They provide a more accelerated curriculum of mathematics, science, and language arts. Most specialize in mathematics and science and some in literature, foreign language, and arts. They all include Grades 1 to 11, are spread throughout the country, and are almost all in urban areas. A very small percentage of the sample attended "education centers," a public school type found only in Moscow, serving certain neighborhoods but not necessarily selective (Table 1).

With our additional teacher questionnaire for ninth-grade teachers, we collected data on three different measures of teacher "quality" found by empirical studies to be related in varying degrees to student achievement and achievement gains: teacher preservice training, teacher experience, and teacher certification categories (for a summary, see Ladd, 2008). Our data show that most ninth-grade teachers (64%) in students' mathematics classrooms in our sample received their mathematics preservice training in faculties of education rather than university mathematics departments (17%). The other 19% received their degrees in other fields, mostly science. Most teachers had substantial experience teaching mathematics—an average of 22 years—and had taught the sample students for an average of 3.5 years, or since the sixth grade.

Our third measure of teacher quality, Russian teacher certification category, is specific to Russian education, but other types of teacher certification in the United States have been found to have significant, albeit small, effects on student achievement (Boyd et al., 2006; Clotfelter et al., 2007; Harris & Sass, 2009). One feature of the certification process in Russia is that both principal evaluations of the teacher's teaching and the quality of the teacher's students' academic work are taken into account. An additional condition is that certification usually takes place once during a five-year period and a teacher with the second highest category qualification has to wait at least two years before she can apply for the highest category. Thus, teachers who have achieved the higher categories usually have considerably more work experience, but there is variation in the work experience of higher category teachers. Because of this nonautomatic teacher professional grading system, the Russian education data provide at least some measure of teacher teaching skills beyond work experience. Thirty-six percent of the teachers reported that they had achieved an official Russian government–issued "high" category certification, which we redefine for greater clarity as the "highest category" certification; 42% reported that they had achieved a "first" category certification, which we redefine as the "second highest" category certification; 16.6% reported they had achieved a "second" category certification, which we redefine as the "third highest" category certification; and only 5.6% reported a "no category" certification, which we redefine as the "lowest" category certification (Table 1). We also collected information on teachers' teaching workload, which averages 24 hours per week; time spent outside the class on nonteaching tasks, which averages 3.3 hours per week; and time spent in administrative work, which averages 3.4 hours.

For OTL, we employed the three indices of exposure to mathematics concepts defined in the PISA 2012 reports. These indices are (a) exposure to applied mathematics concepts, (b) exposure to work problems, and (c) exposure to formal mathematics concepts, specifically algebra and geometry (OECD, 2013c). The three are defined by PISA researchers in terms of a particular question or as combinations of questions from the student questionnaire. The sample means and *SEs* for the three indices are also shown in Table 1.

Estimation Strategy

Our estimation strategy is intended to reduce the bias in typical estimates that use cross-section international test score data and teacher data that cannot be linked to individual students. The goal is to assess more accurately the impact that improving classroom and school resources have on students' PISA math achievement and the policy recommendations that the OECD has made using their more biased estimates. Because we were able to collect data on students' previous achievement and can identify almost all students with their ninth-grade teachers, we can make less biased estimates than the OECD of teacher and teaching effects on student performance.

At the center of our analysis is a model of how the knowledge students bring from home interacts with school and classroom/teacher factors to produce student learning (Goldstein, Bonnet, & Rocher, 2007; Houtenville & Conway, 2008; Ladd, 2008; Levin, 1980; Rivkin et al., 2005). Our model primarily focuses on the resources that students bring to classrooms, the additional resources they are subject to when they enter classrooms, and how classroom resources in particular impact student mathematics achievement. Especially important for more accurately assessing how school resources affect PISA math outcomes at the end of ninth grade, the model also includes a measure of student math knowledge accumulated at the end of eighth grade.

Student resources in our model include individual student characteristics—baseline TIMSS scores, gender, age, and individual family academic resources, including student reported books in the home and mother's education—and approximations of student class/school composition effects as measured by average family academic resources of students in the class and school, specifically the percentage of students in the class reporting higher than total sample median books in the home and the type of school the students attend—regular or selective. The resources students are subject to in classrooms include teachers' capacity to teach the material as measured by the type of mathematics preservice education they received, their level of teacher certification category, and their years of experience teaching mathematics. Teachers expose students to mathematics concepts (OTL) that influence student learning gains directly and indirectly through the capacity of teachers to teach these concepts—we use three OECD definitions of OTL as measures of this exposure. All exposure data are student reported in the PISA student questionnaire. In addition, we include the distribution of teacher workload as a classroom variable. In the model, the outcome of this process is individual students' mathematics achievement.

Statistical Approach

Education within the classroom takes place through a complex process. In particular, student inputs such as family resources and classroom inputs such as teacher characteristics and OTL are systematically related to each other and to student outcomes. To better understand the direct and indirect impacts of various inputs on student outcomes, it is helpful to model these complex relationships explicitly.

Based on the production function literature, three hypotheses underlie our model (Boyd et al., 2006; Clotfelter et al., 2007; Goldstein et al., 2007; Ladd, 2008; Levin, 1980). First, we hypothesize that teacher category is related to teacher experience and teacher preservice mathematics preparation as well as to the classroom average of students' socioeconomic background. The relationship between teacher category and the classroom average of students' socioeconomic background reflects the notion that students and teachers are not allocated to each other randomly, but partly on the basis of students' family academic resources. These relationships are summarized by Equation 1 as follows:

$$\mathrm{TC_j} = \mathrm{C_1} + \gamma_1 \mathrm{TExp_j} + \sum \gamma_2 \mathrm{TEduc_j} + \gamma_3 \mathrm{AvgX_{ij}} + \mathrm{e_{ij}}, \qquad (1)$$

where TC_j = math teacher *j*'s teacher category; $TExp_j$ = teacher *j*'s years of teaching experience, in years of teaching mathematics; $TEduc_j$ = teacher *j*'s type of preservice mathematics education; $AvgX_{ij}$ = percentage of students in classroom *j* not including student *i* that report books in the home higher than total sample median.

Second, we hypothesize that OTL is related to teacher category, teacher experience, teacher preservice mathematics education, and the classroom average of students' family academic resources. In this formulation, OTL acts as a complex mediator of teacher qualifications, in which teachers who are better at teaching mathematics are more likely to expose students to more difficult formal mathematics. What and how much teachers teach students are further influenced by the academic resources students bring to class. Teachers are probably less likely to expose students with low levels of family resources to a high level of formal mathematics compared to students with high levels of family resources. At the same time, students with low levels of family resources are less likely to have a higher category teacher who is better at teaching mathematics. Equation 2 summarizes these relationships as follows:

$$OTL_{ij}=C_2+\sum\beta_1 TC_j+\beta_2 TExp_j+\sum\gamma_4 TEduc_j+\gamma_5 AvgX_{ij}+e_{ij}, \qquad (2)$$

where OTL_{ij} = exposure to one of three math concepts reported by student i in classroom j. The three math concepts we include as variables are those derived by the OECD from the PISA student questionnaire and used in the OECD's PISA analysis (OECD, 2013c)—exposure to "formal mathematics," exposure to "applied math," and exposure to "word problems."

Student achievement is cumulative and is a function of previous achievement and students' family academic resources. Student achievement is also a function of class- or school-level characteristics such as teacher quality, OTL, the average level of family academic resources among students in the classroom, and school selectivity. Typically, however, students' PISA performance is estimated without controlling for students' previous achievement, so we too estimate such a model (Equation 3). We call this model our "typical PISA cross-section model,"

$$\begin{aligned}A_{ijPISA2012}=C_3+\sum b_1 X_{ij}+b_2 AvgX_{ij}+\sum c_2 TC_j+c_3 TExp_j\\ +\sum c_4 TEduc_j+\sum dTAct_j+\sum fOTL_{ij}+\sum gS_i+e_{ij},\end{aligned} \qquad (3)$$

where $A_{ijPISA2012}$ = standardized (mean = 0, SD = 1) PISA mathematics score (2012) for student i in classroom j; X_{ij} = a vector of family characteristics of student i in classroom j; $TAct_j$ = a vector of teacher j's time allocated to different activities (classes, administration, and out-of-class activities); OTL_{ij} = a vector of the three types of exposure to math; S_i = a vector of school types (regular, gymnasium, lyceum, and education center); and e_{ij} = an error term.

A standard problem inherent in estimating the relation between classroom inputs and student mathematics achievement is that students accumulate mathematics knowledge before schooling and over many years in school. We attempt to address this problem in our model by controlling for students' eighth-grade TIMSS score as well as their family academic resources. Specifically, we estimate the following equation:

$$\begin{aligned}A_{ijPISA2012}=C_4+a_1 A_{ijTIMSS}+\sum b'_1 X_{ij}+b'_2 AvgX_{ij}+\sum c'_2 TC_j+c'_3 TExp_j\\ +\sum c'_4 TEduc_j+\sum d'TAct_j+\sum f'OTL_{ij}+\sum g'S_i+e'_{ij}.\end{aligned} \qquad (4)$$

Equation 4 controls for students' accumulated achievement at the beginning of the "treatment year" (ninth grade). Equation 4 is a "typical value-added model." It estimates less biased relations between school resources and student academic achievement than the "typical cross-section" model.

We estimate six variations of the Equation 4 model to test whether estimates change when conditioning on different combinations of teacher characteristics and OTL. We begin with a regression that includes individual student characteristics and student class/school composition variables—average student books in the home in each class and the type of school the student attends, specifically whether a "regular" school, an "educational center," or one of two types of selective schools—a "gymnasium" or "lyceum." In the second regression, we add the type of teachers' preservice training in mathematics—specifically whether this has been in a university mathematics department, the reference category; in an education school; or whether the teacher has not been trained in mathematics as a specialty—and teachers' experience teaching mathematics and experience squared. Both teachers' preparation in subject matter and teachers' experience have been shown in other studies to have a significant effect on student performance. These studies show that experience tends to be less important beyond 10 years, hence the quadratic component. In the third regression, we add teacher certification category and in the fourth regression, the distribution of the teacher's workload. In regressions four through six, we add each of the three types of mathematics exposure, one at a time, since they are quite highly correlated with each other.

To test whether the estimated relations between student PISA achievement and teacher qualifications and OTL are heterogeneous across groups, we also estimate the model in Equations 4 through 6 for two categorizations of students. The two categorizations are (a) by student family academic resources (low, 0–25; middle, 26–100; and high, >100, levels of books in the home) and (b) by baseline student math achievement, divided into four TIMSS benchmark levels: combined Benchmarks 1+2, since only a small number scored at Benchmark 1, and Benchmarks 3, 4, and 5, where 5 is the highest level.

Due to the correlation of student error terms within as opposed to between schools, we estimate cluster-corrected Huber-White estimators for Equations 1 to 4. This is standard practice in the economics of education literature. In a second set of analyses (results not shown for the sake of brevity), we use a multilevel (random effects) model that separates the individual student characteristics from the class and school characteristics. Our results and associated conclusions are substantively the same.

Challenges in Identifying the Model Parameters

To identify the parameters of our model, we face two main challenges. The first challenge is that of selection bias. Selection bias can result from the nonrandom assignment of teachers and students across schools or across

Table 2
Distribution of Teachers by Category, Students' TIMSS Scores, and Family Academic Resources (percentage)

	TIMSS Benchmarks					
Teacher category	1	2	3	4	5	Total[a]
Highest	12.8	22	34.3	42.6	53.4	36.6
Second highest	63.4	52.3	41.7	35.5	31	41
Third highest	17	18.6	17.8	15.8	12	16.4
Lowest	6.8	7.1	6.2	6.1	3.5	6

	Family academic resource groups			
Teacher category	0–25 BIH	26–200 BIH	> 200 BIH	Total[a]
Highest	29.3	37.8	43.5	36.6
Second highest	45.4	40.1	36.9	41
Third highest	19.9	15.8	13.3	16.4
Lowest	5.4	6.3	6.3	6

Source. Russia PISA-TIMSS Survey, 2011–2012.
Note. TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment; BIH = books in the home.
[a]Total percentages of teacher categories in the two parts of the table are slightly different because of missing values in books in the home, because teacher categories come from ninth grade (PISA teacher questionnaire), and both TIMSS benchmarks and books in the home come from the TIMSS survey.

classrooms within schools. Higher achieving or greater family academic resource students may be assigned to teachers of higher quality. Principals may likewise assign teachers to students on the basis of teacher quality (Rothstein, 2009). If teachers and students are nonrandomly assigned across and within schools, as suggested by the estimates in Table 2, the coefficients of achievement gain we estimate for teacher characteristics and OTL may be overestimated. Controlling for students' baseline (TIMSS 2011) in our value-added model can reduce selection bias but may not eliminate it (Raudenbush, 2004; Rubin, Stuart, & Zanutto, 2004).

We further attempt to reduce the bias arising from the nonrandom assignment of students across classrooms/schools by controlling for the average family academic resources of students in each classroom and for the school type the student attends. Both average family resources in the classroom and school type may be good proxies for family motivational differences even within groups of families with similar academic resources. More motivated parents within a group of families with similar academic resources or with similarly low or high scoring students are more likely to try to place their children in classrooms with higher family academic resources or send their children to more selective schools. Controlling for these two variables should remove some selection bias of assignment to better teachers inherent in classroom and school selection.

Some analysts argue that controlling for the average family resources students bring to class underestimates the contribution schooling makes to student performance since better resourced students and their families raise teacher expectations and the level of subject matter that teachers can teach their students (OECD, 2013b). Although this is likely true, it ignores the selection process in which families of students with more academic resources are able to place their children into classrooms/schools with more highly qualified teachers, known to offer a more advanced curriculum, and known to have students with higher levels of academic resources. Attributing the higher performance of students in these classrooms/schools either to better teaching or OTL is an overestimate of the effects of school resources (OECD, 2013b).

The second challenge to identifying the model parameters is that the questions in the PISA survey available to measure OTL—exposure to formal mathematics concepts, exposure to applied mathematics, and exposure to word problems—do not ask students to specify *when* they were exposed to these concepts and types of problems. Thus, we cannot be sure that the OTL in the model is specifically a ninth-grade "treatment."

We are helped in dealing with this challenge by the peculiarities of the Russian educational system. More than 80% of the students in our sample were in the same classroom and with the same teacher in both eighth and ninth grades. Thus, exposure can be related to the ninth-grade teacher whether it took place in eighth or ninth grade. In addition, the concepts covered by the PISA questions on OTL are associated with eighth- and ninth-grade math curricula. Thus, a student who reports more exposure to algebra and geometry (the PISA formal math variable) probably got that greater exposure because he or she was with one particular teacher that exposed the student to those concepts. We do not know whether that took place in the ninth grade; yet, because we control for the eighth-grade TIMSS score, we can argue that the estimated coefficients of these OTL variables measure their effect on PISA outcomes above and beyond students' eighth-grade math performance.

Besides these two challenges, TIMSS and PISA differ in their objectives and the kinds of skills they measure. Although the content areas of the two math tests overlap, TIMSS math tasks address subject mastery level by the eighth grade as defined by standard school math curricula that are consistent with Russia's national mathematics curriculum. PISA math tasks, on the other hand, are designed to assess how well 15-year-olds that are still in school apply skills to practical, real-life situations and problems (Dossey, McCrone, O'Sullivan, & Gonzalez, 2007; Gronmo & Olsen, 2006). Many of the more difficult PISA mathematics tasks require considerable reading and the interpretation of reading distractors to determine the precise mathematics problem to solve. Such tasks test skills that are generally not taught in Russian schools, so that when we measure value-added mathematics gains using the PISA instrument as the posttest, it could be that teacher qualifications may be less identified with gains than had a TIMSS-type instrument been used as the posttest. But this difference in test objectives should not bias our estimated parameters of the relation of teacher characteristics and OTL to students' PISA performance, since we are fundamentally interested in how much these schooling inputs influence PISA performance, controlling for past mathematics performance.

Results

Teacher Qualifications, OTL, and Students' Family Academic Resources

Our estimates of Equations 1 and 2 support the arguments that measures of teacher quality are correlated, OTL is related to teacher quality, and both teacher quality and OTL are related to the average family academic resources in the class. These are important in shaping how we estimate and interpret estimates of the relation between teacher quality and student achievement.

Estimates from Equation 1 confirm two of our hypotheses. First, one of our measures of teacher "quality"—a teacher's category in the Russian government's teacher rating system—is related to other measures of teacher quality, implying that we need to be concerned with correlation among our measures of teacher quality. For example, teacher category is positively and significantly related to teacher preservice preparation in math and teacher experience. Teachers with preservice mathematics in education programs or no formal preservice preparation in mathematics are 2.3 and 2.9 times more likely to be highest category teachers than teachers with university mathematics degrees and also more likely to be either highest or second highest category teachers. In addition, teacher category is also related to average family academic resources in the class, implying that we need to be concerned with selection bias in identifying teacher quality effects on achievement (Table 3). The relationships between having a highest category teacher or either a highest or second highest category teacher in Grade 9 and the average family academic resources in the class are positive and large (Column 1, Table 3).

Estimates from Equation 2 also support our hypothesis that OTL is related to some measures of teacher quality and family academic resources, reinforcing the notion that exposure to mathematics concepts is not randomly distributed in classrooms. The estimates also show that this relationship varies somewhat by type of OTL (Table 4).

Table 3

Estimated Likelihood of Student Having Highest or Second Highest Category Teacher, Related to Teacher and Class Characteristics, Ninth-Grade Class, 2012

	Highest Category Classroom Teacher	Highest or Second Highest Category Classroom Teacher
Teacher's preservice math in education/ pedagogy program[a]	2.30* (1.16)	1.47 (0.77)
Teacher's preservice not in math or math education	2.92* (1.67)	3.50* (2.28)
Teacher's experience teaching subject	1.11* (0.06)	1.15*** (0.06)
Teacher experience squared	1.00 (0.00)	1.00 (0.00)
Class mean student books in the home (% > sample median BIH)	1.73*** (0.26)	1.48** (0.29)
Constant	0.07*** (0.05)	0.32* (0.21)
Observations	4,389	4,389

Source. Russia PISA-TIMSS Survey, 2011–2102.
Note. Robust standard errors in parentheses. TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment; BIH = books in the home.
[a]Reference variable for teacher education is preservice mathematics preparation in university mathematics program.
*p < .10. **p < .05. ***p < .01.

In sum, measures of teacher quality and OTL are related, and as recognized in OECD reports (OECD, 2013a), educational systems do not distribute qualified teachers or OTL equally across classrooms. Rather, groups of students with more family academic resources are more likely to have more qualified mathematics teachers, greater exposure to formal mathematics concepts, and less exposure to applied mathematics concepts. The findings suggest that without controls for student class/school composition, we would misestimate the relationships between teacher quality, OTL, and student achievement.

Estimating PISA Mathematics Achievement

Our "typical PISA cross-section model" (Equation 3) replicates the findings in PISA reports that greater exposure to qualified teachers (OECD, 2010) and OTL (OECD, 2013a, 2013c) can contribute significantly to higher PISA achievement. Note that unlike the OECD estimates, we use data on teachers linked to students. More specifically, the results show that in addition to the typically large positive relation between PISA mathematics score and various individual student family resource measures as well as student class/school composition effects, PISA mathematics achievement is related to teacher

Table 4
Students' Exposure to Mathematics Concepts (OTL) Related to Ninth-Grade Teacher and Class Characteristics, 2012

	Experience With Applied Math	Exposure to Word Problems	Familiarity With Formal Mathematics
Highest category teacher	−0.06 (0.06)	−0.03 (0.07)	0.06 (0.08)
Second highest category teacher	−0.18** (0.08)	−0.16** (0.08)	0.02 (0.09)
Lowest category teacher	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)
Teacher's preservice math in education/pedagogy	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Teacher preservice no formal math education	0.21*** (0.08)	0.12* (0.07)	0.19*** (0.07)
Teacher's years of experience in subject	0.15** (0.07)	0.09 (0.07)	0.09 (0.07)
Teacher experience squared	0.08 (0.11)	0.07 (0.10)	0.03 (0.12)
Class mean student BIH (% > sample median BIH)	−0.10*** (0.02)	−0.00 (0.02)	0.13*** (0.03)
Constant	0.09 (0.10)	0.05 (0.11)	−0.08 (0.10)
Observations	2,908	2,901	2,920
Adjusted R^2	0.014	0.004	0.024

Source. Russia PISA-TIMSS Survey, 2011-2012.
Note. Robust standard errors in parentheses. Reference variables: teacher category = third highest; teacher preservice = university degree in mathematics program. OTL = opportunity to learn; BIH = books in the home; TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment;
$^{**}p < .05$. $^{***}p < .01$.

preservice education in mathematics but not to other measures of teacher capacity, such as years teaching mathematics or highest category teachers (Table 5, Columns 1–5). The coefficient of the relationship between PISA achievement and preservice mathematics training in education programs rather than in university mathematics programs is large, ranging from −.16 to −.21. The estimate is statistically significant at the 10% or 5% level, depending on the model. Students with teachers who had no formal mathematics degree or mathematics education degree—they usually received a degree in science or science education—also scored lower but not significantly. As noted, PISA achievement is not significantly related to teachers' experience in teaching mathematics, which has been identified as a causal factor affecting student achievement in the United States (Ladd, 2008). Yet, counterintuitively, PISA achievement is positively related to having a teacher who spends more hours in administrative tasks.

The results from our "typical cross-section" model also show that PISA achievement is positively and significantly related to various measures of

Table 5
Estimated Student Achievement, PISA 2012

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Student age (eighth grade)	−0.18***	−0.18***	−0.18***	−0.15***	−0.16***	−0.14***
Female	−0.10***	−0.10***	−0.11***	−0.09**	−0.09**	−0.09**
Books in home 11–25	0.12	0.11	0.11	0.23	0.19	0.18
Books in home 26–100	0.28**	0.28**	0.27**	0.36***	0.33**	0.30**
Books in home 101–200	0.38***	0.38***	0.37***	0.48***	0.45***	0.41***
Books in home 200+	0.39***	0.38***	0.38***	0.47***	0.42***	0.38***
Mother's education < HS	−0.04	−0.04	−0.03	0.02	0.01	−0.01
Mother's education postsecondary	0.27***	0.28***	0.28***	0.30***	0.29***	0.27***
Mother's education university	0.40***	0.40***	0.40***	0.37***	0.37***	0.36***
Mother's education graduate school	0.70***	0.69***	0.67***	0.61***	0.61***	0.60***
Mother's education missing	0.05	0.05	0.06	0.06	0.05	0.06
Class average BIH (% > sample median)	0.17***	0.16***	0.16***	0.15***	0.16***	0.15***
School type: gymnasium	0.33***	0.31***	0.28**	0.25**	0.25**	0.24**
School type: lyceum	0.52***	0.55***	0.49***	0.47***	0.47***	0.44**
School type: educational center	−0.14	−0.11	−0.13	−0.23	−0.21	−0.17
Teacher preservice math in education/pedagogy		−0.16*	−0.18**	−0.20**	−0.20**	−0.21**
Teacher preservice no formal math education		−0.17	−0.19	−0.25*	−0.23*	−0.24*
Years teaching math		0.02	0.01	0.01	0.01	0.01
Years teaching math squared		−0.00	−0.00	−0.00	−0.00	−0.00
Teacher highest category			0.05	0.06	0.03	0.01
Teacher second highest category			−0.05	−0.05	−0.07	−0.08
Teacher lowest category			−0.24	−0.28	−0.27	−0.29
Workload classes				−0.00	−0.00	0.00
Workload out of classes				0.00	0.00	0.00

(continued)

Table 5 **(continued)**

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Workload administration				0.01**	0.01**	0.01**
Exposure applied math (z-score)				−0.14***		
Exposure word problems (z-score)					0.04**	
Exposure formal math (z-score)						0.15***
Constant	2.09***	2.14***	2.19***	1.74**	1.86**	1.67**
Observations	4,389	4,389	4,389	2,908	2,901	2,920
Adjusted R^2	0.191	0.197	0.201	0.219	0.202	0.224

Source. Russia TIMSS-PISA sample, 2011-2012.
Note. Reference variables: 0–10 books in the home; mother's education = high school complete; teacher preservice education = degree in mathematics; teacher third highest category; school type = regular secondary school. Standard errors of coefficient estimates available on request. HS = high school; BIH = books in home; TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment.
*$p < .10$. **$p < .05$. ***$p < .01$.

OTL (Table 5, Columns 5–7). In estimating the regressions, we converted the three OTL variable scales shown in Table 1 to standardized scores with a mean value of zero and an *SD* = 1. The estimated coefficients therefore show that a 1 *SD* increase in exposure to formal mathematics is associated with a .15 *SD* increase in students' PISA scores. A 1 *SD* increase in exposure to word problems is associated with a .04 *SD* increase in PISA scores and a 1 *SD decrease* in exposure to applied math with a .14 *SD* increase in PISA scores.[2]

Thus, the PISA reports may be correct that some teacher characteristics and some types of OTL are associated with higher student PISA scores. However, failing to control for students' previous achievement may result in over- or misestimating classroom factors that contribute positively to student outcomes as "value added." Our estimates in the following show that this is indeed the case.

Estimating PISA "Value Added" Relative to Students' TIMSS Performance

When we control for students' previous achievement (eighth-grade TIMSS scores) in our "typical value-added model" (Equation 4), the various relationships of PISA to classroom variables are weaker than for the PISA estimates without controlling for students' TIMSS scores. First, the negative coefficient of preservice training in education (pedagogy) departments ranges from −.14 to −.15, smaller than in the PISA cross-section estimate (Table 6, Columns 2–6). The magnitude of the coefficient of preservice non–math education is also smaller and generally not significant. Second, the coefficients of teacher categories relative to third lowest teacher category continue to be not statistically significant. Third, the coefficient of teacher administrative workload is neither positive nor significant. And fourth, the coefficient for formal math exposure remains positive (.09) and significantly related to PISA achievement (Column 6), albeit much smaller than in the cross-section model. The coefficient for applied math exposure remains negative (–.07) and significantly related to PISA achievement (Table 6, Column 4) but also much smaller than in the cross-section model. The coefficient of OTL in the form of more exposure to word problems is not significant in the typical value-added model (Table 6, Column 5). The continued positive relation between exposure to formal mathematics and PISA math scores in ninth grade when we control for student TIMSS scores suggests that the effect of such OTL exposure persists even when we include a measure designed to pick up the effects of such exposure in eighth grade and earlier.

All these results support the notion that increasing (a) the proportion of teachers with preservice training in university mathematics departments and (b) OTL in the form of increased exposure to formal mathematics would contribute to higher Russian student achievement on the PISA. However, as expected, all the coefficients of these variables are smaller than the size

Table 6
Estimated Student Achievement, PISA 2012, Including TIMSS 2011 Math Score

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
TIMSS math score 2011	0.53***	0.53***	0.53***	0.52***	0.53***	0.52***
Female	−0.08***	−0.08***	−0.08***	−0.06*	−0.06*	−0.06*
Class Average BIH (% > sample median)	0.09***	0.08***	0.08***	0.08**	0.08**	0.08**
School type: gymnasium	0.18**	0.16**	0.15*	0.14*	0.14*	0.13*
School type: lyceum	0.17*	0.19*	0.19*	0.21*	0.21*	0.19*
School type: educational center	−0.21	−0.18	−0.20	−0.31**	−0.30*	−0.27*
Teacher preservice math education		−0.14**	−0.15**	−0.15*	−0.15*	−0.15**
Teacher preservice no math education		−0.17	−0.17	−0.21*	−0.20	−0.20
Teacher years teaching math		0.01	0.00	0.00	0.00	0.00
Years teaching math squared		−0.00	−0.00	−0.00	−0.00	−0.00
Teacher highest category			0.00	0.00	−0.02	−0.03
Teacher second highest category			0.04	0.04	0.03	0.02
Teacher lowest category			−0.17	−0.22	−0.21	−0.23
Workload classes				0.00	0.00	0.00
Workload out of classes				0.00	0.00	0.00
Workload administration				0.00	0.00	−0.00
Exposure applied math (*z*-score)				−0.07***		
Exposure word problems (*z*-score)					0.01	
Exposure formal math (*z*-score)						0.09***
Constant	1.60**	1.69**	1.68**	1.17	1.22	1.11
Control for student FAR	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,389	4,389	4,389	2,908	2,901	2,920
Adjusted R^2	0.437	0.440	0.442	0.437	0.431	0.441

Source. Russia TIMSS-PISA sample, 2011–2012.
Note. Reference variables: 0–10 books in the home; mother's education = high school complete; teacher preservice education = degree in mathematics; teacher third highest category; school type = regular secondary school. Standard errors of coefficient estimates available on request. HS = high school; BIH = books in home; TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment; FAR = family academic resources.
$*p < .10$. $**p < .05$. $***p < .01$.

of the "typical PISA cross-section" estimates—much smaller in the case of the OTL variables.

Estimating PISA "Value Added" Relative to Students' TIMSS Performance for Students From Low, Middle, and High Family Academic Resource Groups

Dividing our analysis of the PISA "typical value-added" model into three student family academic resource groups—lower (0–25 books in the home), middle (26–100 books in the home), and higher (>100 books in the home)—we find that the relation of our measures of teacher quality and OTL to student PISA scores varies by groups (for reasons of space, we only present the final three of our stepwise regressions). According to Table 7 (Columns 1–9), preservice training in mathematics taken in education programs is negatively related to PISA scores for all three groups of students, but it is smaller and not significant for the lowest family academic resource group. The coefficient of lowest teacher category relative to the third highest teacher category is negative for all three groups, but it is not significant in any group.

Furthermore, whereas exposure to applied mathematics is negatively and significantly related to PISA scores in all three groups, the negative impact is larger for students with lower family academic resources than for students with middle family academic resources and much larger than for students with higher family academic resources. Similarly, exposure to formal mathematics is large, positive, and statistically significant for students with lower and middle family academic resources but not significant for students with higher family academic resources. Again, counterintuitively, students in the highest family academic resource group with teachers that spend more hours in outside-of-class activities score significantly higher on PISA.

Estimating PISA "Value Added" Relative to Students' TIMSS Performance for Students Scoring at the Five TIMSS Benchmark Levels

The estimates of PISA achievement across groups of students achieving different levels of TIMSS benchmarks in eighth grade, controlling for eighth-grade TIMSS score, show that the coefficients of PISA scores estimated for teacher characteristics are different for students scoring at lower TIMSS benchmark levels from those scoring at the highest benchmark level (Table 8). In benchmark groups 1+2 combined, students with lowest category teachers are associated with significantly lower PISA scores compared to students with teachers in the third lowest certification category, the reference group. The effect size is large, about .4 to .5 standard deviations. In the highest benchmark group, it is students with the highest category teachers that have higher PISA scores, but these coefficients are not significant. There is also a negative relation in a higher benchmark group (4) of having

Table 7

Estimated Student Achievement, PISA 2012, by Student FAR Level, Controlling for TIMSS Math Score

	Students in Lowest FAR (0–25 BIH)			Students in Middle FAR (26–100 BIH)			Students in Highest FAR (>100 BIH)		
	Model 3	Model 5	Model 6	Model 3	Model 5	Model 6	Model 3	Model 5	Model 6
TIMSS math score 2011	0.44***	0.46***	0.44***	0.57***	0.57***	0.57***	0.59***	0.59***	0.59***
Female	−0.08	−0.07	−0.08*	−0.06	−0.06	−0.05	−0.03	−0.04	−0.04
Class average BIH (% > sample median BIH)	0.05	0.06	0.06	0.05	0.05	0.04	0.09**	0.09***	0.09**
Teacher preservice math education	−0.10	−0.09	−0.11	−0.16*	−0.17*	−0.16*	−0.16**	−0.15**	−0.16**
Teacher preservice no math education	−0.19	−0.18	−0.19	−0.22	−0.22*	−0.21	−0.23*	−0.22*	−0.23*
Teacher years teaching math	−0.00	−0.00	−0.00	0.01	0.01	0.01	−0.00	−0.00	−0.00
Years teaching math squared	0.00	0.00	0.00	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
Teacher highest category	−0.10	−0.12	−0.15	−0.03	−0.04	−0.05	0.15	0.14	0.12
Teacher second highest category	−0.00	−0.01	−0.03	−0.04	−0.05	−0.05	0.16	0.16	0.14
Teacher lowest category	−0.31	−0.31	−0.32	−0.26	−0.22	−0.26	−0.05	−0.05	−0.07
Workload classes	0.00	0.00	0.00	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
Workload out of classes	0.00	0.00	0.00	−0.01	−0.01	−0.01	0.02**	0.02**	0.02*
Workload administration	−0.01	−0.01	−0.01	0.00	0.00	0.00	0.00	0.00	0.00
Exposure applied math	−0.11***			−0.07**			−0.05**		
Exposure word problems		−0.03			0.04			0.02	
Exposure formal math			0.10***			0.11***			0.05
Constant	1.03	1.10	1.20	2.28*	2.11	1.77	0.91	1.01	0.88
Control for individual student FAR	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for school type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	897	893	900	1,053	1,050	1,058	958	958	962
Adjusted R^2	0.310	0.297	0.312	0.447	0.441	0.453	0.491	0.489	0.492

Source. Russia TIMSS-PISA sample, 2011–2012.

Note. Reference variables: 0–10 books in the home; mother's education = high school complete; teacher preservice education = degree in mathematics; teacher third highest category; school type = regular secondary school. Standard errors of coefficient estimates available on request. Student FAR controls are student age, books in the home, and mother's education. School types are gymnasium, lyceum, and education center, with regular secondary school, the reference variable. HS = high school; BIH = books in home; TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment; FAR = family academic resources.

* $p < .10$. ** $p < .05$. *** $p < .01$.

Table 8

Estimated Student Achievement, PISA 2012, Controlling for TIMSS Math Score, by TIMSS Benchmarks

	TIMSS Benchmark 1+2			TIMSS Benchmark 3			TIMSS Benchmark 4			TIMSS Benchmark 5		
	(1) Model 4	(2) Model 5	(3) Model 6	(4) Model 4	(5) Model 5	(6) Model 6	(7) Model 4	(8) Model 5	(9) Model 6	(10) Model 4	(11) Model 5	(12) Model 6
TIMSS math score 2011	0.35***	0.35***	0.34***	0.31***	0.31***	0.31***	0.31***	0.32***	0.32***	0.37***	0.38***	0.37***
Female	−0.11*	−0.12*	−0.11*	−0.02	−0.01	−0.02	−0.01	−0.01	−0.02	−0.18	−0.19	−0.19
Class average BIH (% > sample median BIH)	−0.05	−0.06	−0.06	0.04	0.05	0.03	0.04	0.05	0.04	0.18***	0.19***	0.18***
Teacher preservice math in education/pedagogy	−0.18	−0.20	−0.18	−0.13	−0.13	−0.15	−0.14*	−0.13*	−0.14*	−0.11	−0.09	−0.09
Teacher preservice no formal math education	−0.39	−0.41	−0.37	−0.14	−0.14	−0.17	−0.22*	−0.21	−0.22*	−0.15	−0.09	−0.13
Years teaching math	−0.02	−0.02	−0.02	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.01
Years teaching math squared	0.00	0.00	0.00	0.00	0.00	0.00	−0.00	−0.00	−0.00	−0.00	−0.00	−0.00
Teacher highest category	−0.09	−0.08	−0.10	0.01	0.00	−0.03	0.01	−0.01	−0.01	0.21	0.14	0.15
Teacher second highest category	−0.13	−0.13	−0.13	−0.00	−0.00	−0.04	0.16	0.16	0.15	0.19	0.12	0.11
Teacher lowest category	−0.47*	−0.42*	−0.48*	−0.38	−0.38	−0.38	0.04	0.05	0.04	−0.04	−0.05	−0.07
Workload classes	0.00	0.00	0.00	0.00	0.00	0.00	−0.00	−0.00	−0.00	−0.01	−0.01	−0.01
Workload out of classes	0.00	0.00	0.00	0.01	0.01	0.01	0.03**	0.03**	0.03**	−0.02*	−0.02**	−0.02**
Workload administration	0.00	0.00	0.00	−0.01**	−0.01	−0.01*	0.00	0.00	0.00	−0.00	−0.00	−0.00
Exposure applied math	−0.02			−0.06*			−0.07***			−0.11***		
Exposure word problems		0.00			−0.03			0.05**			0.04	
Exposure formal math			0.05			0.10***			0.10**			0.18***
Constant	0.97	1.05	0.96	2.87**	2.86**	2.64**	0.72	0.83	0.78	1.08	1.32	0.96
Control for student FAR	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for school type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	580	576	581	921	917	925	996	997	1,001	411	411	413
Adjusted R^2	0.179	0.182	0.184	0.122	0.116	0.136	0.147	0.141	0.149	0.328	0.313	0.331

Source. Russia TIMSS-PISA sample, 2011–2012.

Note. Reference variables: 0–10 books in the home; mother's education = high school complete; teacher preservice education = degree in mathematics; teacher third highest category; school type = regular secondary school. Standard errors of coefficient estimates available on request. Student FAR controls are student age, books in the home, and mother's education. School types are gymnasium, lyceum, and education center, with regular secondary school, the reference variable. HS = high school; BIH = books in home; TIMSS = Trends in International Mathematics and Science Survey; PISA = Program for International Student Assessment; FAR = family academic resources.

$^{*}p < .10$. $^{**}p < .05$. $^{***}p < .01$.

a teacher trained in math in an education program rather than in a university mathematics program. The effect size is about −.13. These results suggest that teacher "quality" is positively related to PISA achievement but not consistently or systematically across groups with different levels of "initial" academic achievement.

Students' PISA scores are also positively related to exposure to formal mathematics in the middle and higher benchmark groups (3, 4, and 5) and negatively related to exposure to applied math problems in all but the lowest two TIMSS benchmark groups (Table 8). The absence of a significant relation of exposure to formal mathematics content or applied math for students with relatively low levels of initial mathematics achievement suggests that increasing the opportunity to learn more formal mathematics or decreasing the OTL of applied mathematics may not increase PISA scores across all mathematics ability groups. These two components of OTL, particularly formal mathematics, seem to have a much stronger relation to PISA for students with high initial mathematics achievement score than for students with middle-level initial mathematics achievement.

Discussion and Conclusions

The many recommendations for educational improvement generated by international agencies such as the OECD are based on analyses of cross-section international tests. We argue that these analyses produce potentially biased results because they incorrectly attribute all the knowledge students gain over the course of their schooling to the resources of their current school/grade and, in the case of PISA, are unable to identify students with teachers, so generally attribute the performance of each student to the average of teacher resources in their current school. We found that these two problems, particularly the first, tend to overestimate the effects of teacher resources and opportunity to learn indicators on student performance claimed in OECD documents.

We used unique data from a random sample of eighth-grade Russian students who initially took the TIMSS 2011 mathematics test in the eighth grade and then the PISA 2012 mathematics test in the ninth grade. We had access to the data that TIMSS gathered on their eighth-grade classes/teachers and schools, and we used follow-up data on their ninth-grade classes/teachers and schools. This longitudinal data set allowed us to measure the "gains" that students make in their ninth-grade year in one country and link classroom factors to those gains. Although still not entirely free from selection bias, our value-added results are considerably more precise than the results presented in international assessment reports that seek to identify education policies to improve mathematics achievement.

The main reason for the greater precision in our results is that we have a baseline test taken by the students in our sample a year earlier, at the end of eighth grade. However, our analysis also pays more systematic attention than earlier studies of international assessments to the importance of students' family academic resources in students' PISA achievement. The study accounts for the influence of students' family academic resources on their test gains in three ways: (a) by controlling for family resources in estimating gains on the PISA test; (b) by controlling for the fact that students are in schools and classrooms with peers with similar family academic resources—such composition effects are positively and significantly related to PISA gains; and (c) by estimating the relationships between student achievement gains and classroom resources for subgroups of students with different levels of family academic resources and achievement.

These empirical findings from Russia support the logic that "better" mathematics preparation for mathematics teachers and more exposure for students to formal mathematics have positive, significant effects on student PISA mathematics performance. But they also suggest that OECD claims about raising students' PISA scores by improving school/classroom resources are overstated.

We find that these effects vary across students with different levels of family academic resources and students with different levels of math knowledge accumulated by the end of eighth grade. This should caution policymakers against assuming that the same teacher "improvements" and OTL policies would have similar impacts across the entire student population.

Our results also do not lend support to the idea that PISA scores for Russia's lowest family resource and achievement students can be improved merely by increasing teacher quality, although they do suggest that lower math ability students are benefited greatly by not having the lowest category teachers. Our results suggest that the positive effect on PISA scores of teachers with stronger math preparation are consistent for students who are in the middle to higher groups of family academic resources and those who score in the broad middle to high-middle range of TIMSS benchmark levels. Students who come from families with lower academic resources or those who score at the lower and middle TIMSS benchmark levels appear less likely to benefit from teachers with "better" mathematics training. Thus, if the objective is to equalize learning gains by focusing on improving the academic performance of low family academic resources or of least "math able" students, putting them with "better math prepared" teachers may not work.

Our finding that Russian students with initially lower levels of TIMSS scores facing third highest category teachers appear to make significantly larger gains in ninth grade than students with lowest category teachers needs to be interpreted cautiously since only 6% of teachers have the lowest category. The positive relation of having a highest category teacher to student achievement gains on the PISA is limited to students scoring at Benchmark 5, and even at that benchmark level, the estimated effect is not statistically significant. The results also suggest that the "logical" policy implicit in

PISA recommendations of assigning higher quality teachers to lower math achieving and lower family academic resource students is unlikely to improve those students' mathematics performance. It could be that those higher quality teachers are more suited to teaching more advanced mathematics to students with higher math skills.

We need to be careful in drawing this conclusion for another reason. Lower scoring eighth-grade students may appear to be doing better with third highest category teachers because their more motivated parents have been successful in avoiding having their children assigned to classrooms or schools with "lowest category" math teachers and high scoring TIMSS students may be making larger gains from eighth to ninth grade in such classrooms because they have highly motivated parents who made sure they were in classrooms with highest category teachers who are perceived, or even known, to make large gains in math. In both cases, students from more highly motivated families would have made these larger gains in ninth grade even if they had not been with second or highest teachers, depending on the benchmark level. Although we do control for average family academic resources in the classroom and for type of school, it is possible that even with these controls, we are not picking up differential parent motivation across regular middle/secondary schools.

Two of the PISA OTL mathematics exposure indicators—exposure to formal mathematics (algebra and geometry) and exposure to applied mathematics—are, for all students together, significantly related to students' PISA scores in our "typical value-added model" estimates—positively for formal mathematics and negatively for applied math. Increasing OTL through more exposure to formal mathematics appears to have a relatively large potentially positive impact on students with low family academic resources but does not offer much promise for increasing the PISA scores of students scoring lower on the TIMSS test. This suggests that more exposure to formal mathematics most benefits lower family resource students with higher mathematics ability but not the most "disadvantaged" group in education—those students who come from lower resource families and are not able in mathematics. Nevertheless, the result that exposing students with low family academic resources but with middle and higher initial TIMSS scores to more formal mathematics is related to higher PISA scores is important since, on average, lower FAR students are much less likely to get high exposure to formal mathematics (Table 4). This is a much more nuanced result than the policy conclusion in PISA reports that exposing all lower FAR students to more school resources will help them make larger gains.

Another (counterintuitive) finding in our results is that PISA scores are not significantly related to student exposure to word problems. These results are particularly surprising because more exposure to test items that require greater reading skills (word problems) should help students do better on the PISA, which often uses such items.

To conclude, this study serves as a cautionary tale. It is not a good idea to use cross-sectional international test results, such as the PISA findings, to make sweeping generalizations about what works in education. Our results suggest that there *are* ways that improving teacher education and increasing the opportunity for students to learn formal mathematics can raise student achievement, and some of these are consistent with PISA claims. But our study also shows that if policymakers are to invest effort and money in such reforms, they should have much more precise, less biased estimates than what cross-sectional international and national results can provide. Whereas "big" international studies such as the TIMSS or PISA are useful in identifying broad trends, there is still no substitute for careful causal inference analysis carried out in particular social contexts, such as in one country's or one region's low-income or low-scoring schools, in order to determine what works in those contexts to improve student learning.

Notes

The data used in this study came from the Russian panel study "Trajectories in Education and Careers" (TrEC – http://trec.hse.ru/). The authors gratefully acknowledge financial support from the Basic Research Program of the National Research University Higher School of Economics and supported within the framework of a subsidy by the Russian Academic Excellence Project "5-100."

[1]The books in the home (BIH) variable we use to estimate student class composition is highly correlated with a class composition variable using average mother's education. Our regression results are substantially similar when we employ individual or class-aggregated measure of relative mother's education as a control variable rather than BIH.

[2]The negative and significant effect of applied mathematics on students' Program for International Student Assessment (PISA) gains in Russia accords with the Russian results in the PISA report, but Russian results do not accord with the overall finding for applied mathematics in the PISA 2012 report (no control for previous mathematics achievement). The overall finding suggests a quadratic relation between such exposure and PISA mathematics performance (OECD, 2013a).

References

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y-M. (2010). Teachers' mathematical knowledge, teachers' cognitive activations in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, *1*(2), 176–216.

Carnoy, M., Chisholm, L., & Chilisa, B. (Eds.). 2012. *The low achievement trap*. Pretoria, South Africa: Human Sciences Research Council Press.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673–682.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Darling-Hammond, L. (2009). *Educational opportunity and alternative certification: New evidence and new questions*. Palo Alto, CA: Stanford University: SCOPE Policy Brief.

Dee, T. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, *43*(3), 528–554.

Dossey, J. A., McCrone, S., O'Sullivan, C., & Gonzalez, P. (2007). *Problem solving in the 2003 PISA and TIMSS assessments, technical report*. Washington, DC: Institute of Educational Studies.

Fuchs, T., & Woessmann, L. (2004). *What accounts for international differences in student performance? A re-examination using PISA data* (IZA Discussion Paper 1287). Munich, Germany: Institute for the Study of Labor.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Education Evaluation and Policy Analysis*, *22*(2), 129–145.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*(3), 252–286.

Gronmo, L. S., & Olsen, R. V. (2006, November). *TIMSS versus PISA: The case of pure and applied mathematics*. Paper presented at the 2nd IEA International Research Conference, Washington, DC.

Hanushek, E. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, *24*(3), 1141–1177.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7), 798–812.

Harris, D. N., & Sass, T. R. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management*, *28*(1), 55–80.

Hill, H., Rowan, B., & Ball, D. (2005). Effects of teachers' mathematics knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Houtenville, A. J., & Conway, K. S. (2008). Parental effort, school resources, and student achievement. *Journal of Human Resources*, *43*(2), 437–453.

Kukla-Acevedo, S. (2009). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. *Economics of Education Review*, *28*(1), 49–57.

Ladd, H. (2008). Teacher effects: What do we know? In G. Duncan & J. Spillane (Eds.), *Teacher quality: Broadening the debate* (pp. 3–26). Evanston, IL: Northwestern University.

Lee, V. (2000). Response: Opportunities for design changes. In D. Grissmer & J. Ross (Eds.), *Analytic issues in the study of student achievement* (pp. 237–248). Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement.

Levin, H. M. (1980). Educational production theory and teacher inputs. In C. Bidwell & D. Windham (Eds.), *The analysis of educational productivity, Vol. 2: Issues in microanalysis*. Cambridge, MA: Ballinger.

Loveless, T. (2014). *Lessons from the PISA-Shanghai controversy*. Retrieved from http://www.brookings.edu/research/reports/2014/03/18-pisa-shanghai-loveless.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

OECD. (2010). *PISA 2009 results: What students know and can do (Vol. I)*. Paris: Author.

OECD. (2011). *Lessons from PISA for the United States: Strong performers and successful reformers in education*. Paris: Author.

OECD. (2013a). *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed (Vol. II)*. Paris: Author.

OECD. (2013b). *PISA 2012 results: What makes schools successful (Vol. IV)*. Paris: Author.

OECD. (2013c). *PISA 2012 results: What students know and can do (Vol. I)*. Paris: Author.

Raudenbush, S. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121–129.

Rivkin, S., Hanushek, E. A., & Kain, J. A. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rubin, D., Stuart, E. A., & Zanutto, E. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Schleicher, A. (2014, April). *Why care about international comparisons? Evaluating school systems to improve education*. Presented at the American Educational Research Association, Philadelphia, PA.

Schmidt, W. H., McKnight, C., Houang, R., Wang, H., Wiley, D., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33.

Van Klaveren, C. (2011). Lecturing style teaching and student performance. *Economics of Education Review*, *30*(4), 729–739.

White, K. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*(3), 461–481.

Woessmann, L., Luedemann, E., Schuetz, G., & West, M. (2009). *School accountability, autonomy and choice around the world*. London: Edward Elgar.

Manuscript received August 24, 2014
Final revision received April 27, 2016
Accepted May 12, 2016