Contents lists available at ScienceDirect

International Journal of Educational Development

journal homepage: www.elsevier.com/locate/ijedudev

Are teachers accurate in predicting their students' performance on high stakes' exams? The case of Russia

Andrey Zakharov[a], Martin Carnoy[*,b]

[a] National Research University Higher School of Economics, Ul. Myasnitskaya 20, Moscow, Russia
[b] Stanford University, School of Education, 485 Lasuen Mall, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 21 September 2014
Received in revised form 23 March 2015
Accepted 27 April 2015

Keywords:
Educational policy
International education
Teacher expectations
High stakes exams

ABSTRACT

The paper focuses on how accurate teachers may or may not be in gauging their class'academic abilities. We use a sample of classrooms in three Russian regions to identify sources of mathematics and Russian teachers' inaccuracies in predicting their high school classes' scores on Russian and mathematics high stakes college entrance tests (the Unified State Exam, or USE). We test the hypothesis that teachers' perceptions of their relationship with their classes are good predictors of such inaccuracies. This is important because teachers often focus on their relationship with the class as an end in itself or as a means to engaging students. Good teacher–student relations may indeed result in more students' learning, but perhaps not nearly as much as teachers' believe. We find that both Russian and mathematics teachers make inaccurate predictions of their class' high stakes examination results based on how they perceive their relationship with their class. Teachers who believe they have a very good relationship with the class significantly overestimate their class' performance on the USE, and those who perceive a poor relationship, underestimate their class' performance, although this underestimate is generally not statistically significant.

1. Introduction

Much of the literature on effective teaching emphasizes the ideal of teachers as reflective professionals capable of individualized approaches to student learning (for example, Cohen, 1993; Darling Hammond, 1996). An implicit assumption underpinning this ideal is that teachers are accurate, fair judges of their students' abilities, and that they can (and should) individualize broad curricular guidelines to fit each student's capacity and learning style.

Two strands of research have questioned this assumption. One strand, going back to the 1960s, argues that teachers may not be neutral observers of students' abilities, that teachers' expectations may vary among students, and that teachers' expectations (positive or negative) can affect students' performance (Rosenthal and Jacobson, 1968). More recently, this research has turned to finding students' (and teachers') characteristics that may affect teacher expectations in particular subjects, and, hence, their students' performance (Brophy, 1983; Rosenthal and Rubin, 1978; Rosenthal, 1994). A number of studies find teacher gender bias—teachers viewing boys as having greater math and science skills and girls as having greater literary skills (Qing, 1999; Ready and Wright, 2011; Riegle-Crumb and Humphries, 2012; Shepardson and Pizzini, 1992), but others find no evidence of teacher gender bias (Dusek and Joseph, 1983; Madon et al., 1998). Similarly, many studies have found teacher race/ethnic bias (Ready and Wright, 2011; Rubie-Davies et al., 2006; Tenenbaum and Ruck, 2007), and social class bias (Auwarter and Aruguete, 2008; Ready and Wright, 2011). This may also relate to how teachers view students in different academic tracks (Kelly and Carbonaro, 2012; Oakes, 1985; Page, 1987; Tach and Farkas, 2006). Teacher gender and ethnicity have also been shown to play a role in affecting the performance of students' of particular gender and ethnicity (Dee, 2005; McKown and Weinstein, 2008; Ready and Wright, 2011; Van den Bergh et al., 2010). Some of these studies estimate causal effects and show that teachers' subjective judgment—consciously or unconsciously—can and does affect students' academic outcomes.

Although much less studied, the second strand of research argues that teachers' expectations for students' performance compared to actual results may differ not because of conscious or unconscious "biases," but because of what Jussim and Harber (2005) called "predictive validity without self-fulfilling influence." Ferguson (2003) argued that both teacher inaccuracy and bias

[*] Corresponding author. Tel.: +1 6508567722.
E-mail addresses: ab.zakharov@gmail.com (A. Zakharov), carnoy@stanford.edu (M. Carnoy).

http://dx.doi.org/10.1016/j.ijedudev.2015.04.007
0738-0593/© 2015 Elsevier Ltd. All rights reserved.

regarding student performance are deviations from a ‘true” benchmark, namely how much students have actually learned and their performance on measures of their learning. Teachers may be poor predictors of student performance because teachers may misestimate how much certain teaching practices or classroom conditions positively or negatively affect student learning and test performance. Teachers’ expectations based on these views of practices and classroom conditions can be inaccurate but not necessarily biased if the difference of predicted and actual class average test scores does not vary systematically according to classroom (students’) characteristics.

In this paper, we focus on this second strand of “predictive validity,” namely how accurate teachers may or may not be in gauging their class’ academic abilities. We use a sample of classrooms in three Russian regions to identify sources of mathematics and Russian teachers’ inaccuracies in predicting their high school classes’ scores on Russian and mathematics high stakes college entrance tests (the Unified State Exam, or USE). We test the hypothesis that teachers’ perceptions of their relationship with their classes are good predictors of such inaccuracies. This is important because teachers often focus on their relationship with the class as an end in itself or as a means to engaging students. Good teacher–student relations may indeed result in more students’ learning, but perhaps not nearly as much as teachers’ believe.

Teachers should be able to predict how their class will score on either the mathematics or Russian section of the USE, since in the 11th grade of Russian schools great emphasis is placed on preparing for this examination, including homework assignments and practice tests. Teachers in Russia usually teach the same students for two years in either mathematics or Russian. We control for the characteristics of a class that could “bias” teachers’ expectations of students’ performance. We also control for two variables that should help teachers make better predictions of their class’ performance on the USE: (a) students’ grades in the first semester of the 11th grade (information about students’ previous performance) and (b) teachers’ years of teaching experience (accumulated expertise).

We find that both Russian and mathematics teachers make inaccurate predictions of their class’ high stakes examination results based on how they perceive their relationship with their class. Teachers who believe they have a very good relationship with the class significantly overestimate their class’ performance on the USE, and those who perceive a poor relationship, underestimate their class’ performance, although this underestimate is generally not statistically significant. Teachers’ view of their relationship with the students in their class is not significantly related to the gender composition of their class in either math or Russian, but Russian teachers are significantly more likely to report a very good relation with their class when the cultural capital (Bourdieu and Passeron, 1976; Bourdieu, 1986)[1] of their students is higher, even controlling for students’ academic performance. This may suggest some “cultural capital bias” on the part of Russian teachers.

The paper is structured as follows: in the next section, we describe our study’s Russian secondary education context; Section 3 describes our methodology and data; Section 4, the results, and the final section discusses the results and concludes.

2. The advantages of studying the Russian case

Russian high school is a particularly interesting context in which to study teachers’ accuracy in assessing how well their students have learned an academic subject. High school teachers generally know their students well in Russia. Almost all students have the same mathematics teacher and the same Russian language teacher for both high school grades (10–11th). In many instances, they even have the same teacher for each subject since the 5th grade. Thus 11th grade teachers have known their students for at least two and often more years. Further, for demographic reasons and student attrition (around 40% of students go to vocational school after the 9th grade) there is usually one math and one Russian class in the 11th grade of high school and it is usually of a smaller size than classes in middle or primary schools. All these factors help teachers to become more familiar with their students’ attitudes and abilities. Teacher expectations for their students should therefore be rather accurate in high school.

As some analysts have noted, studies of teacher perceptions rarely use objective measures of students’ outcomes; therefore bias in teacher expectations may be reported even if it didn’t take place (Ready and Wright, 2011). The data we use for outcomes are the students’ Unified State Examination (USE) scores. The USE is a high school exit/college entrance test used throughout Russia since 2009. It has some distinct advantages as a measure of students’ actual performance. It is a standardized test graded by agencies external to the school and thus provides an “objective” measure of students’ achievement. It is curriculum based: it measures students’ performance on the subjects they had studied. Finally it is high stakes test. It is required for graduation from upper secondary school and also serves as an entry exam to all universities in the country. Russian language and mathematics are of special importance as they are mandatory subjects to sit in USE. Many university departments require USE results in mathematics, and all departments in all universities require the results in Russian. Almost all 11th grade students want to attend university, and the USE results determine their choices. High stakes tests such as the USE are therefore likely to serve as better measures of students’ actual performance because of the their high motivation to perform well (Bishop, 1997).

3. Data

3.1. Survey timing and sampling

We use data from a sociological survey conducted in May 2010 in three Russia regions: Pskovskaya and Yaroslavskaya *oblasts* and Krasnoyarsky *krai*. These regions were selected because they provide significantly different demographic, social and economic contexts for high school education. Yaroslavskaya *oblast* is a small region in the center of Russia, rather average in regional ratings of social and economic development. Pskovskaya *oblast* is located in the northwest part of the country. It is also small in terms of population and relatively poor economically. Finally Krasnoyarsky *krai* is a large region in the east of Russia (in Siberia) that is one of the nation’s most highly developed regional economies.

The survey was applied toward the end of a school year, two to four weeks before the USE examination, in May. By that time, teachers had a great deal of information about students’ performance, behavior and family background, considerably increasing the likelihood of accurate teacher predictions of student performance on the USE.

The data were collected based on a stratified random sample that represented final year students in each selected region. Using a list of all schools, we grouped them into strata. Among the parameters for stratification were the type of settlement (rural, urban, regional center), school type (regular school, school with advanced study of some subjects, *gymnasiums*, and *lyceums*), and

[1] Bourdieu and Passeron (1976) defined cultural capital rather broadly on the one hand as parents' education and the intellectual climate at home, and, on the other, rather narrowly as the quality of language and verbal interaction at home. We used a proxy for cultural capital—more than 100 books in the home as reported by the student. We also collected data on students' parents' education, but a large fraction of students do not answer this question, and for those students that did, parents' education was highly correlated with books in the home. For a discussion of books in the home as a proxy for family academic resources, see Carnoy and Rothstein (2013).

Table 1
Descriptive statistics of variables used in the analysis.
Source: Russian Regions School Sample, 2010.

	Mathematics classes			Russian classes		
Variables	N	Mean	SD	N	Mean	SD
Student characteristics						
USE score	2938	45.36	14.75	2938	61.04	11.33
Class characteristics						
Actual class average USE score	182	44.35	8.92	182	59.97	6.62
Teacher predicted class average USE score	169	40.12	12.88	173	50.06	12.51
Difference of actual and teacher predicted class average USE scores	169	4.32	12.11	173	9.90	10.45
Teacher reported neutral or bad relationship with class (0 = no, 1 = yes)	182	0.14	0.35	182	0.13	0.34
Teacher reported good relationship with class (0 = no, 1 = yes)	182	0.59	0.49	182	0.50	0.50
Teacher reported very good relationship with class (0 = no, 1 = yes)	182	0.26	0.44	182	0.37	0.48
% students reported very good relationship with teacher	182	0.28	0.21	182	0.30	0.23
Class size	182	16.93	7.83	182	17.02	7.94
Advanced subject study (0 = no, 1 = yes)	182	0.34	0.47	182	0.19	0.40
Class mean first semester grades in Algebra/Russian	182	3.62	0.32	182	3.67	0.28
% girls in class	182	0.58	0.19	182	0.58	0.19
% students with >100 books in home	182	0.42	0.21	182	0.42	0.21
Teacher experience: ≦10 years (0 = no, 1 = yes)	182	0.10	0.30	182	0.08	0.28
Teacher experience: 11–20 years (0 = no, 1 = yes)	182	0.22	0.42	182	0.23	0.42
Teacher experience: 21–30 years (0 = no, 1 = yes)	182	0.37	0.49	182	0.48	0.50
Teacher experience: ≧31 years (0 = no, 1 = yes)	182	0.31	0.46	182	0.21	0.41
Teacher has 2nd or no category (0 = no, 1 = yes)	182	0.16	0.37	182	0.15	0.36
Teacher has 1st category (0 = no, 1 = yes)	182	0.48	0.50	182	0.42	0.50
Teacher has highest category (0 = no, 1 = yes)	182	0.36	0.48	182	0.42	0.50

high school size (number of students in the 11th grade). In each stratum, schools were selected for the survey with a simple random sampling procedure. In total we sampled 39 schools in Pskovskaya *oblast,* 42 schools in Yaroslavskaya *oblast* and 46 schools in Krasnoyarsky *krai*. In sampled schools, all the 11th grade students were selected for the survey.

We also surveyed their mathematics and Russian language teachers and school principals. In total, 2938 students and mathematics and Russian teachers in 182 classrooms participated in the survey. We surveyed 805 students and teachers in 53 classes in Pskovskaya *oblast,* 986 students and teachers of 60 classes in

Table 2
Estimated teachers' views on relationship with their class, by teacher view of class (ordered probit estimates).
Source: Russian Regions Schools Sample, 2010.

	Mathematics classes					Russian classes				
Variables	(1) Full probit	(2) Neutral/ bad	(3) Good	(4) Very good	(5) Full probit	(6) Full probit	(7) Neutral/ bad	(8) Good	(9) Very good	(10) Full probit
% Students reporting very good relationship with teacher	1.51***	−0.30***	−0.17**	0.47***		0.87**	−0.15**	−0.17*	0.32**	
	(0.47)	(0.10)	(0.08)	(0.15)		(0.42)	(0.07)	(0.09)	(0.16)	
Class size	−0.02*	0.00*	0.00*	−0.01*	−0.03**	−0.03***	0.01**	0.01***	−0.01***	−0.03***
	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
Advanced subject study (0 = no, 1 = yes)	0.05	−0.01	−0.01	0.01	0.18	0.13	−0.02	−0.03	0.05	0.14
	(0.18)	(0.04)	(0.02)	(0.06)	(0.18)	(0.21)	(0.03)	(0.05)	(0.08)	(0.21)
Class mean first semester grades in Algebra/Russian	0.38	−0.07	−0.04	0.12	0.43	1.17***	−0.20***	−0.23***	0.43***	1.29***
	(0.35)	(0.07)	(0.05)	(0.11)	(0.34)	(0.34)	(0.06)	(0.09)	(0.13)	(0.33)
% Girls in class	−0.12	0.02	0.01	−0.04	−0.06	0.26	−0.04	−0.05	0.09	0.22
	(0.56)	(0.11)	(0.07)	(0.17)	(0.55)	(0.54)	(0.09)	(0.11)	(0.20)	(0.53)
% Students with >100 books in home	0.47	−0.09	−0.05	0.15	0.47	0.79*	−0.13*	−0.16	0.29*	0.89*
	(0.48)	(0.10)	(0.06)	(0.15)	(0.45)	(0.47)	(0.08)	(0.10)	(0.17)	(0.47)
Teacher experience: ≦10 years (0 = no, 1 = yes)	−0.62	0.16	−0.00	−0.16**	−0.72*	−0.27	0.05	0.04	−0.09	−0.30
	(0.38)	(0.12)	(0.05)	(0.08)	(0.38)	(0.33)	(0.07)	(0.04)	(0.11)	(0.33)
Teacher experience: 21–30 years (0 = no, 1 = yes)	−0.10	0.02	0.01	−0.03	−0.22	−0.51**	0.09**	0.10**	−0.19**	−0.51**
	(0.26)	(0.05)	(0.03)	(0.08)	(0.25)	(0.21)	(0.04)	(0.04)	(0.08)	(0.21)
Teacher experience: ≧31 years (0 = no, 1 = yes)	0.04	−0.01	−0.00	0.01	−0.06	−0.23	0.04	0.04	−0.08	−0.28
	(0.28)	(0.06)	(0.03)	(0.09)	(0.27)	(0.26)	(0.05)	(0.04)	(0.09)	(0.25)
Teacher has 2nd or no category (0 = no, 1 = yes)	−0.02	0.00	0.00	−0.01	−0.10	−0.34	0.07	0.05*	−0.12	−0.34
	(0.28)	(0.06)	(0.03)	(0.09)	(0.28)	(0.24)	(0.06)	(0.03)	(0.08)	(0.25)
Teacher has highest category (0 = no, 1 = yes)	0.21	−0.04	−0.03	0.07	0.24	0.34*	−0.06*	−0.07	0.13*	0.37*
	(0.20)	(0.04)	(0.03)	(0.06)	(0.19)	(0.20)	(0.03)	(0.05)	(0.08)	(0.20)
Cut1: constant	0.31				0.06	2.99***				3.19***
	(1.24)				(1.22)	(1.12)				(1.13)
Cut2: constant	2.20*				1.87	4.69***				4.86***
	(1.25)				(1.22)	(1.14)				(1.15)
Observations	182	182	182	182	182	182	182	182	182	182

Notes: columns (1, 6): full probit regression estimate; columns (2, 7): probit marginals for teachers with neutral/bad view of their relation with their class; columns (3, 8): probit marginals for teachers with good view of their relation; columns (4, 9): probit marginals for teachers with very good view of their relation; columns (5, 10): full probit regression estimate not including students' view of their relationship with their teacher.

Table 3
Structural equation model estimates for mathematics teachers' predicted class average mathematics USE score and class average actual mathematics USE score (*z*-score). Source: Russian Regions Schools Sample, 2010.

	Without covariates			With covariates		
Variables	Direct effect (1)	Indirect effect (2)	Total effect (3)	Direct effect (4)	Indirect effect (5)	Total effect (6)
Math teachers' predicted class average math USE (*z*-score)						
Math teacher relationship with class is neutral or bad (0 = no, 1 = yes)	−0.36*			−0.41**		
	(0.22)			(0.21)		
Math teacher relationship with class is very good (0 = no, 1 = yes)	0.73***			0.59***		
	(0.18)			(0.17)		
% Students reporting very good relationship with math teacher	−0.24			−0.62*		
	(0.36)			(0.36)		
Advanced math study: 6–8 h/week (0 = no, 1 = yes)				0.34**		
				(0.16)		
Class size				−0.01		
				(0.01)		
% Girls in class				−0.45		
				(0.40)		
% Students with more than 100 books in home				0.73**		
				(0.32)		
Class mean 1st semester final mark in Algebra				0.73***		
				(0.23)		
Math teacher experience: ≦10 years (0 = no, 1 = yes)				0.23		
				(0.23)		
Math teacher experience: 21–30 years (0 = no, 1 = yes)				−0.34		
				(0.23)		
Math teacher experience: ≧31 years (0 = no, 1 = yes)				−0.24		
				(0.25)		
Math teacher has 2nd or no category (0 = no, 1 = yes)				0.10		
				(0.18)		
Math teacher has highest category (0 = no, 1 = yes)				0.28*		
				(0.15)		
Constant	−0.07			−2.54**		
	(0.13)			(1.09)		
Variance (e.avPredMathz)	0.87***			0.66***		
	(0.13)			(0.12)		
Class average actual math USE score (*z*-score)						
Teacher predicted class average math USE (*z*-score)	0.42***		0.42***	0.17**		0.17**
	(0.12)		(0.12)	(0.08)		(0.08)
Math teacher relationship with class is neutral or bad (0 = no, 1 = yes)	0.04	−0.15	−0.11	−0.04	−0.07	−0.11
	(0.14)	(0.10)	(0.16)	(0.16)	(0.04)	(0.17)
Math teacher relationship with class is very good (0 = no, 1 = yes)	0.04	0.31	0.35	0.08	0.10*	0.17
	(0.19)	(0.11)	(0.19)	(0.15)	(0.05)	(0.14)
% Students reported very good relationship with math teacher	0.93**	−0.10	0.83	0.36	−0.10	0.25
	(0.43)	(0.14)	(0.48)	(0.34)	(0.06)	(0.34)
Advanced math study: 6–8 h/week (0 = no, 1 = yes)				0.48***	0.06	0.54***
				(0.13)	(0.04)	(0.13)
Class size				0.01	−0.00	0.01
				(0.01)	(0.00)	(0.01)
% Girls in class				−0.47	−0.08	−0.55
				(0.42)	(0.07)	(0.44)
% Students with more than 100 books in home				0.87**	0.12	0.99**
				(0.36)	(0.09)	(0.38)
Class mean 1st semester final mark in Algebra				1.02***	0.12	1.14***
				(0.24)	(0.08)	(0.22)
Math teacher experience: ≦10 years (0 = no, 1 = yes)				−0.30	0.04	−0.26
				(0.26)	(0.05)	(0.25)
Math teacher experience: 21–30 years (0 = no, 1 = yes)				0.00	−0.06	−0.05
				(0.16)	(0.04)	(0.17)
Math teacher experience: ≧31 years (0 = no, 1 = yes)				−0.23	−0.04	−0.27*
				(0.15)	(0.04)	(0.16)
Math teacher has 2nd or no category (0 = no, 1 = yes)				−0.07	0.02	−0.05
				(0.26)	(0.03)	(0.26)
Math teacher has highest category (0 = no, 1 = yes)				0.06	0.05	0.11
				(0.12)	(0.03)	(0.12)
Constant	−0.26*			−4.13***		
	(0.15)			(0.93)		
Variance (e.avMathz)	0.78***			0.50***		
	(0.13)			(0.08)		

Notes: robust standard errors in parentheses.
* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

Yaroslavskaya *oblast* and 1147 students and teachers of 69 classes in Krasnoyarsky *krai*.

3.2. Measures

The survey provides rich information about students and their families, teachers, and school curriculum. Students reported their demographic characteristics (sex, age), family background (family structure, books in home, items in home etc.), 11th grade first semester grades and information about their relationship with mathematics and Russian teachers. Teachers reported their demographic and professional characteristics (sex, age, education, experience etc.) and their relationship with their class. They were also asked to predict their class' average USE results in terms of the percentage of students in the class who would perform at different levels on the mathematics or Russian test. Principals provided information about school type, school size and schools' curriculum. Students' characteristics and teachers' and schools' characteristics were merged with students' official USE results in both mathematics and Russian. Regional Ministries/ Departments of education provided the actual USE scores for each student.

The USE is graded on a 0–100 points scale. For comparison purposes, we standardized both subjects' USE scores (mean = 0, SD = 1). We also computed teacher estimates of their prediction of class performance on the USE in their subject. Each teacher was asked to predict the ratio of students in a class who would pass the exam with 0–20, 21–40, 41–60, 61–80 and 81–100 scores. Based on the teachers' answers we estimated the weighted teacher predicted average USE scores in Russian (Russian teacher) and mathematics (mathematics teacher) for each class,[2] and compared these with the actual class mean USE scores in the two subjects, subtracting the weighted teacher predicted class average score from the actual class average USE score.[3]

To measure our variable of interest—the teacher's view of her relationship with her class—both Russian and mathematics teachers were asked to characterize their relationship with the class using a 5-points Likert scale ("very bad," "rather bad," "neutral," "rather good," or "very good"). Taking into account the distribution of teacher answers, we generated a series of dummy variables indicating whether teachers see their relationship with the class as "very good", "good" and "neutral or bad."

We measured a number of class and teacher characteristics in order to control for them as potential sources of teacher bias. Among the class characteristics, we used students' reported views of the relationship with their teacher in each subject (based on students answers we calculated the ratio of students who estimated this relationship as very good), class size, the ratio of girls in the class, and the ratio of students with more than 100 books in home. To control for tracking we created dummy variables with 1 denoted advanced study of the subject (3 or more hours a week for Russian, 6–8 h a week for mathematics) and with 0 denoted study of the subject at basic level (1–2 h a week for Russian, 4–5 h a week for mathematics).

We also controlled for the class previous performance in mathematics and Russian. Students reported their previous achievement (11th grade first semester grades in algebra and Russian). These marks were assigned by teachers at the end of December and reflected the first half-year performance in a subject.[4] Based on these answers we calculated class mean 1st semester grade in Algebra/ Russian.

We measured two main indicators of teacher "skill": teacher experience and teacher qualification category. Teacher experience was measured in years of teaching. We created a series of dummy variables indicating 0–10, 11–20, 21–30, 31 or more years of experience for both Russian and mathematics teachers. Teacher category is assigned as a result of certification process. In 2010, teachers could have "no category" or might be certified with "the second" (the lowest), "the first" or the "highest category" (only a small percentage of teachers in our sample had "no category," so we combined them with "second category" in creating the teacher certification dummy variables). Since such certification takes place over prescribed periods of time for each teacher, teacher category is correlated with teacher experience. However, teacher category provides additional information about teacher skills, since teachers also have to present proof of good performance to attain higher categories.[5] Further, since these are only used as controls, we make no attempt to interpret their separate influence on our dependent variables.

Finally, we measured several school characteristics—location of the school (urban/rural, size of town), school size and school type (those with and without advanced tracks). Since these variables are highly correlated with class characteristics, however, we didn't use them in the analysis.

Descriptive statistics for all variables used in the analysis are shown in Table 1.

4. Estimation strategy

To test our hypothesis that teachers' view of their relationship with the class influences the accuracy of their predictions of student performance, we employ a variant of a structural equation model (SEM). The underlying argument of the model is that teachers' view of their relationship to the class is fashioned by student and teacher characteristics (see Eq. (1)), and that both teachers' predictions of class average performance and actual class average performance on the USE are influenced by teachers' view of their relationship to the class (see Eqs. (2) and (3)). When we include teachers' predicted test score in the estimate for the class' actual USE test score, the coefficient should represent the influence of expectations "net" of the information the teacher has about students in the class' previous performance.

Eq. (1) is estimated as an ordinal probit, with the three teacher views ("very good," "good," and "neutral or bad") as the dependent variable, R_{cs}. Eqs. (2) and (3) are estimated as a SEM using maximum likelihood.[6]

In addition to the SEM, we estimate Eq. (4)—the difference between actual class average USE score and teacher predictions using OLS.

[2] The weighting was done by multiplying the midpoint of each range of scores by the ratio of students the teacher indicated would score at each level.

[3] Teacher predictions for individual students' USE results were not provided by our data. Thus we were not able to take into account the fact that teacher perception of students' performance may vary within the same classroom.

[4] Such grades are a categorical 4-points scale with "2" as the minimum and "5" as maximum. It could be argued that these grades reflect not only students' previous outcomes but teachers' perception of these outcomes (Jussim, 1991). Thus, these grades themselves may reflect content evaluation bias and bias due to teachers' misperception of students' achievements or learning behavior. Our data do not allow us to examine these sorts of biases.

[5] We collected data on other teacher characteristics: gender and level and field of pre-service education. However, there is little variance in these characteristics among Russian teachers. They are almost all women who have completed higher education, most with a major in education (pedagogy).

[6] We also estimated Eqs. (1),(2), and (3) using the STATA SEM algorithm. This yielded essentially identical coefficients as estimating Eq. (1) separately and Eqs. (2) and (3) using the SEM estimation algorithm. From the standpoint of the reader, presenting the estimates for the teachers' view of the relationship with their class makes the results easier to understand.

The model is specified as follows:

$$R_{cs} = \alpha + \beta_1 C_{cs} + e_{cs}; \tag{1}$$

$$E_{cs} = \alpha' + \beta'_1 R_{cs} + \beta'_2 C_{cs} + e'_{cs}; \tag{2}$$

$$Y_{cs} = \alpha'' + \beta''_1 E_{cs} + \beta''_2 R_{cs} + \beta''_3 C_{cs} + e''_{cs}; \tag{3}$$

$$\Delta_{cs} = \alpha''' + \delta_1 R_{cs} + \delta_2 C_{cs} + e'''_{cs}; \tag{4}$$

where

R_{cs} is the teacher view of her relationship with class c in school s,

C_{cs} is a vector of class c characteristics (including teacher characteristics) in school s,

E_{cs} is weighted teacher mean expected (predicted) USE score in class c and school s,

Y_{cs} = mean actual USE score of class c in school s,

$\Delta_{cs} = E_{cs} - Y_{cs}$,

α, β_n, δ_n = regression coefficients,

e_{cs} is an error term.

Note that in all equations, we analyze the data aggregated at the class level. Two datasets (for mathematics and Russian classes) are analyzed separately.

This structural model provides important insights into the relationships between teacher predictions, teachers' opinions of their relationship with students, and the accuracy of their predictions of students' academic performance. But the relations we estimate are likely subject to biases. The most important of these regards the tendency of test results to "regress to the mean" and how that tendency could be correlated with teachers' self-selecting into higher achieving classrooms. Teachers teaching higher achieving classes are likely to feel more positive about them (we estimate this relation in Eq. (1)), and higher achieving groups of students should tend to regress to the mean on any given test—they should tend to average lower scores than previous performance. Conversely, lower achieving students would tend to score higher than previous performance and be taught by teachers who are less likely to have chosen to teach them and therefore feel less positive about them. Thus, our estimates of the error teachers make regarding students performance is partially due not to their feelings about the class, but to the regression to the mean of students' test scores—the resulting δ_1 is likely to be an overestimate of the "true" relation between teacher views of their relationship with the class and the inaccuracy of their predictions.

5. Results

5.1. Estimating teachers' views of their relationship with their class

Teachers' views of their relationship with their class could depend in part on class performance in the subject taught, in part on teachers' experience and teaching skill (as measured by their "category"), and, if the teacher is biased in her perceptions of students with certain "inherent" characteristics, on the characteristics of the students in the class. We test this model by estimating an ordinal probit regression. The ordinal values of the probit regression are the three possible views the teacher can report of her relation with the class—"very good", "good", and"neutral or bad".

When we include the ratio of students who reported having a very good relationship with their mathematics teacher as a covariate, this variable and class size are the only factors that are significantly related to the view the teacher has of the class. The estimates in Table 2, columns 2–4 show that the higher the ratio of students who reported a very good relationship with their teacher, the less likely the teacher is to report a "neutral/bad"or"good" relationship with her class and the more likely she is to report a "very good" relationship. Finally, the larger the number of students in the class, the less likely the math teacher is to report a "very good" relationship with the class.

Students' and Russian teachers' views of their relationship are also significantly related but not as much as in math classes. A teacher's view of her relationship with her class is significantly related to students' 11th grade first semester grades in Russian (positively), as well as to class size (negatively) and to whether a teacher has many years of experience (negatively) and has a highest category classification (positively) (Table 2, columns 6–9). Most important for our analysis, Russian teachers' views of their relationship with the class are significantly related to students' class average cultural capital, as measured by the ratio of students having more than 100 books in the home.

It appears that Russian teachers are much more likely to report having very good relation with their class when the average cultural capital in their class is higher. When we drop students' view of their teacher from the estimate of the view held by teachers (Table 2, columns 5 and 10), the coefficient of the ratio of students having more than 100 books in their home does not change for Russian teachers and is still not significant for math teachers, suggesting that the possible bias of teachers toward higher cultural capital students is confined to Russian classrooms.

5.2. Teachers' predictions of their class' performance on the USE test, teachers' view of their relationship with class, and actual class performance on the USE test

Now we turn to the SEM estimate of the relationship between teachers' predictions of their class' performance on the USE test, the class' actual performance on the test, and the teachers' view of their relationship with their class. We have shown that a teacher's view of her relation with her class is correlated with the class' average first semester grade in the subject and with the ratio of students who reported a very good relationship with the teacher. Additionally, in Russian classes, there may be teacher bias in favor of students with greater family cultural capital. In turn, we would expect that a teacher's prediction of her class' future outcomes on the USE exam in her subject would be strongly related to important "real" factors that likely influence students' academic performance, such as previous academic achievement (as proxied by students' grades), and by the teacher's years of experience teaching.

5.2.1. Mathematics teachers and classes

Our estimates of the mathematics teacher's predictions for her class' performance on the USE test are related to the teacher's view of her relations with the class even when we control for class characteristics, teacher's experience and category, and the ratio of students' who reported a very good relationship with the teacher (Table 3, columns 1 and 4, and Fig. 1). With controls for class characteristics, such as the class' mean first semester grades in mathematics and the proportion of students reporting more than 100 books in the home, the coefficients for a teacher's reported "very good" perceived view of the relationship with the class decreases in size but is still statistically significant. Compared to teachers who estimated their relationship with class as "good," teachers who reported a "very good" relationship predict that their class' average score on the mathematics USE will be about 0.60 standard deviations higher. Having a "neutral or bad" view of her relationship with the class is significantly related to a teacher's prediction of her class' performance on the math USE, and once we control for teacher and class characteristics, the coefficient of

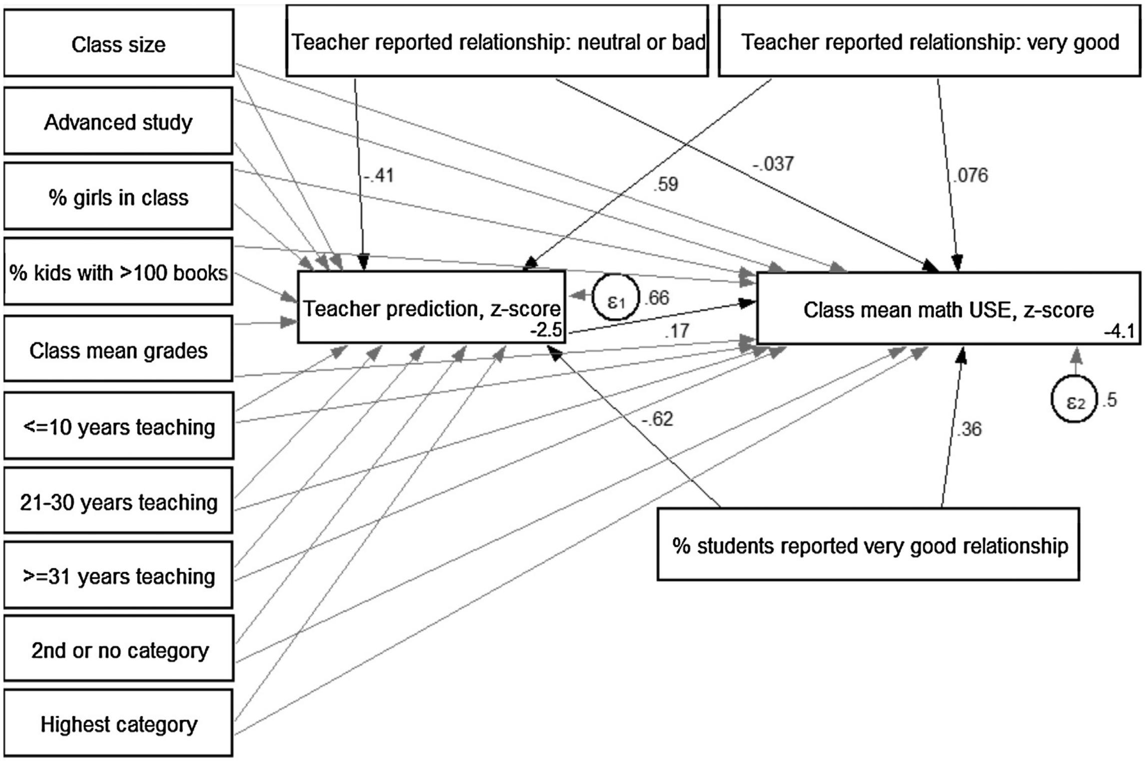

Fig. 1. Path diagram for mathematics teachers' prediction of class average mathematics scores and actual class average mathematics scores on USE examination.
Notes: lines between class student/teacher characteristics and teacher's reported relationship with the class (neutral/bad or very good relative to good)—see Table 2, column 1—have been omitted to make figure easier to read.
Source: Russian Regions Schools Sample, 2010.

having a "neutral or bad" perceived relation with the class becomes larger and more significantly related to lower teacher prediction of class performance.

Table 3 and Fig. 1 also show that the teacher's prediction for class outcomes is significantly related to actual class mean USE for mathematics (see columns 1 and 3, 4 and 6). When we control only for teachers' reported relationship with the class and do not control for other variables, the coefficients of predicted test score on actual test score are large. For mathematics, a one standard deviation increase in teachers' *predicted* class mean USE score is associated with 0.42 standard deviations increase in *actual* class mean score. Yet when we control for class characteristics (including class mean first semester grades in the subject) the coefficient declines to 0.17 standard deviations for mathematics. This suggests that there may be some influence of teacher "expectations" conveyed to the class on how well the class does on the test.

5.2.2. Russian teachers and classes

The estimates of Russian teachers' predictions for class performance on the USE test are somewhat different from those of math teachers. Table 4, (columns 1 and 4), and Fig. 2 show that once we control for class characteristics, teachers' views of whether they have a very good or neutral/bad relationship to the class do not significantly differ in their relation to predicted class results from those teachers who report a good relationship. This suggests that in Russian classes, beyond their knowledge of students' past performance, teachers seem to be less influenced than in mathematics in estimating how well they expect students to do on the USE test by their feelings about their relationship to the students in the class.

One interpretation of this result is that perhaps because language and reading is considered more cognitively "determined," Russian teachers are less convinced than math teachers that their take on the quality of their teaching a particular class—proxied by their view of very good or neutral/bad relations with the class—would influence student outcomes. But more likely, the close relationship in Russian classes between teachers' view of their relationship with students and students' previous academic performance, as well as teachers' "cultural capital" bias, produces the apparent insignificant relation of Russian teachers' view of their class to their expectations for the class on the USE.

We can also conclude that both Russian and math teachers have a "bias" in their exam performance predictions that favors classes with more family cultural capital (books in the home). The robust and positive coefficients for cultural capital controlling for class average first semester grades, provides evidence of such bias. A counter-argument is that these robust coefficients only indicate that in addition to teachers' predications being heavily influenced by their class's previous grades in the subject, teachers are just being realistic in making higher predictions of USE performance for higher cultural capital classes. Tables 3 and 4 present evidence that, even controlling for previous grades, students with higher cultural capital do, indeed, tend to score higher on the USE in both Russian and math.

As in mathematics (Table 3), Table 4 shows that Russian teachers' prediction for class outcomes is significantly related to actual class mean USE for Russian but the effect size is even larger: 0.57 standard deviations when we control only for teachers' reported relationship with the class and do not control for other variables (see Table 4, columns 1 and 3). Yet as in math classes, when we control for class characteristics, the coefficients decline to 0.31 standard deviations for Russian (see Table 4, columns 4 and 6).

A second important result in Table 4 is that Russian teachers' reported views of their relationship with the class are not significantly directly related to actual class performance on the

Table 4
Structural equation model estimates for Russian teachers' predicted class average Russian USE score and class average actual Russian USE score (z-score).
Source: Russian Regions Schools Sample, 2010.

Variables	Without covariates			With covariates		
	Direct effect (1)	Indirect effect (2)	Total effect (3)	Direct effect (4)	Indirect effect (5)	Total effect (6)
Russian teachers' predicted class average Russian USE (z-score)						
Russian teacher relationship with class is neutral or bad (0 = no, 1 = yes)	−0.37			−0.21		
	(0.25)			(0.21)		
Russian teacher relationship with class is very good (0 = no, 1 = yes)	0.40***			0.09		
	(0.15)			(0.13)		
% Students reported very good relationship with Russian teacher	0.95***			0.57**		
	(0.33)			(0.25)		
Advanced Russian study: ≧3 h/week (0 = no, 1 = yes)				0.18		
				(0.16)		
Class size				0.01		
				(0.01)		
% Girls in class				−0.01		
				(0.35)		
% Students with more than 100 books in home				0.80**		
				(0.34)		
Class mean 1st semester final mark in Russian				1.24***		
				(0.23)		
Russian teacher experience: ≦10 years (0 = no, 1 = yes)				0.63**		
				(0.31)		
Russian teacher experience: 21–30 years (0 = no, 1 = yes)				0.14		
				(0.16)		
Russian teacher experience: ≧31 years (0 = no, 1 = yes)				0.24		
				(0.19)		
Russian teacher has 2nd or no category (0 = no, 1 = yes)				−0.26		
				(0.23)		
Russian teacher has highest category (0 = no, 1 = yes)				0.35**		
				(0.15)		
Constant	−0.38**			−5.49***		
	(0.16)			(0.82)		
Variance (e.avPredRussz)	0.85***			0.58***		
	(0.08)			(0.07)		
Class average actual Russian USE score (z-score)						
Teacher predicted class average Russian USE (z-score)	0.57***		0.57***	0.31***		0.31***
	(0.09)		(0.09)	(0.10)		(0.10)
Russian teacher relationship with class is neutral or bad (0 = no, 1 = yes)	0.08	−0.21	−0.13	0.16	−0.06	0.09
	(0.17)	(0.15)	(0.21)	(0.18)	(0.06)	(0.19)
Russian teacher relationship with class is very good (0 = no, 1 = yes)	0.08	0.23**	0.31*	−0.05	0.03	−0.03
	(0.14)	(0.10)	(0.17)	(0.12)	(0.04)	(0.12)
% Students reported very good relationship with Russian teacher	−0.23	0.54***	0.31	−0.40	0.18*	−0.22
	(0.33)	(0.21)	(0.39)	(0.24)	(0.10)	(0.26)
Advanced Russian study: ≧3 h/week (0 = no, 1 = yes)				−0.12	0.06	−0.06
				(0.12)	(0.05)	(0.13)
Class size				0.02*	0.00	0.02**
				(0.01)	(0.00)	(0.01)
% Girls in class				0.47	−0.00	0.47
				(0.35)	(0.11)	(0.37)
% Students with more than 100 books in home				1.20***	0.25*	1.45***
				(0.35)	(0.15)	(0.37)
Class mean 1st semester final mark in Russian				1.02***	0.38***	1.41***
				(0.25)	(0.14)	(0.25)
Russian teacher experience: ≦10 years (0 = no, 1 = yes)				−0.19	0.19**	0.01
				(0.31)	(0.09)	(0.35)
Russian teacher experience: 21–30 years (0 = no, 1 = yes)				0.02	0.04	0.07
				(0.13)	(0.05)	(0.13)
Russian teacher experience: ≧31 years (0 = no, 1 = yes)				−0.00	0.07	0.07
				(0.20)	(0.06)	(0.20)
Russian teacher has 2nd or no category (0 = no, 1 = yes)				−0.24	-0.08	−0.32*
				(0.16)	(0.08)	(0.19)
Russian teacher has highest category (0 = no, 1 = yes)				−0.04	0.11*	0.07
				(0.14)	(0.06)	(0.13)
Constant	0.02			−4.62***		
	(0.13)			(0.90)		
Variance (e.avRussz)	0.71***			0.51***		
	(0.10)			(0.07)		

Notes: robust standard errors in parentheses.
[*] $p < 0.10$.
[**] $p < 0.05$.
[***] $p < 0.01$.

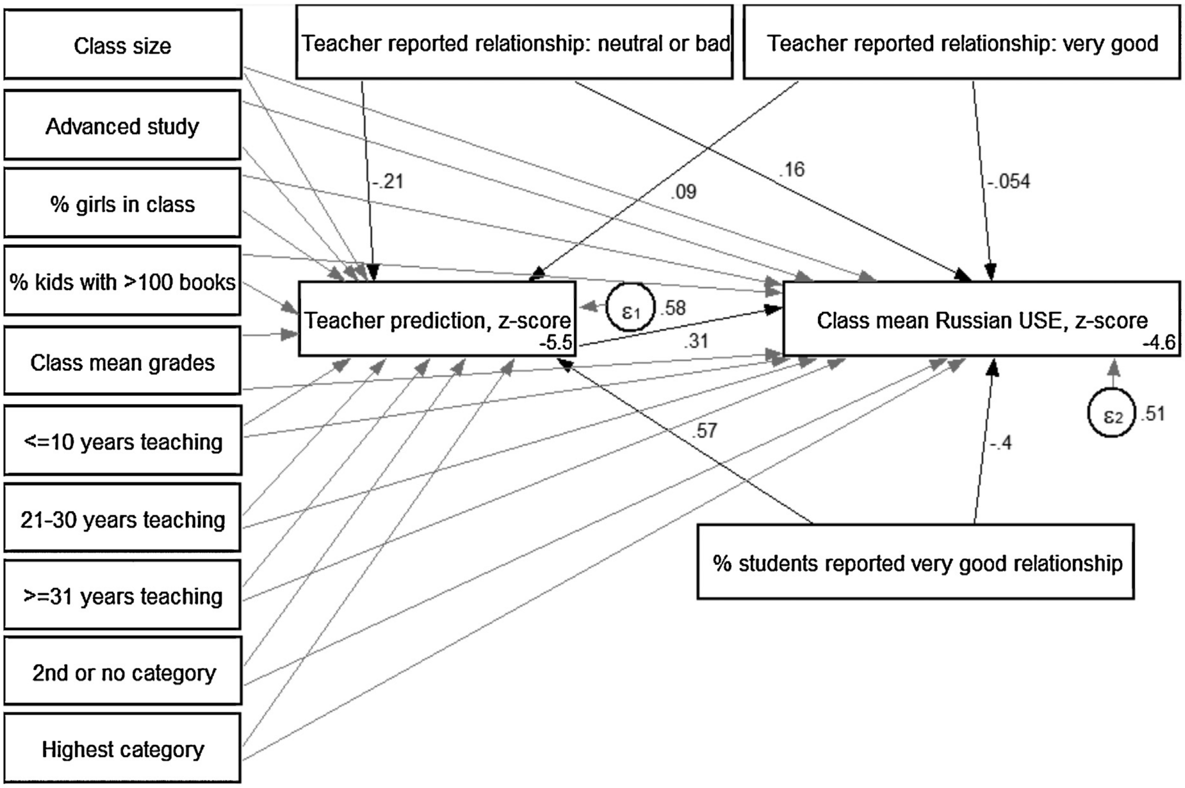

Fig. 2. Path diagram for Russian teachers' prediction of class average Russian scores and actual class average Russian scores on USE examination.
Notes: lines between class student/teacher characteristics/% students reported very good relationship with teacher and teacher's reported relationship with the class (neutral/bad or very good relative to good)—see Table 2, columns 6—have been omitted to make figure easier to read.
Source: Russian Regions Schools Sample, 2010.

USE once we control for class characteristics (columns 4 and 6). Thus, given the class mean first semester grades and ratio of students with more than 100 books in the home—both apparently good predictors of students' test performance— teachers' views of their class relation do not seem to be directly related to actual USE results. But teachers' views of class relations do influence teachers' prediction of class performance on the USE (and teachers' predictions are related to actual scores). This suggests that Russian teachers' perceived very good or neutral/bad relations with their class may influence how well students actually perform on the

Table 5
Estimated difference of actual and teachers' predicted class average USE scores (OLS estimates).
Source: Russian Regions Schools Sample, 2010.

	Mathematics classes				Russian classes			
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher reported neutral or bad relationship with class (0 = no, 1 = yes)	0.25	0.25	0.33	0.36	0.41	0.28	0.32	0.30
	(0.21)	(0.24)	(0.23)	(0.23)	(0.25)	(0.25)	(0.23)	(0.22)
Teacher reported very good relationship with class (0 = no, 1 = yes)	−0.41**	−0.39**	−0.40**	−0.50**	−0.38**	−0.27	−0.18	−0.12
	(0.18)	(0.19)	(0.19)	(0.20)	(0.15)	(0.17)	(0.17)	(0.17)
% students reported very good relationship with teacher				0.85*				−0.82***
				(0.45)				(0.30)
Class size		0.01	0.01	0.01		−0.00	0.01	0.00
		(0.01)	(0.01)	(0.01)		(0.01)	(0.01)	(0.01)
Advanced subject study (0 = no, 1 = yes)		0.11	0.12	0.03		−0.34*	−0.28	−0.25
		(0.20)	(0.19)	(0.21)		(0.18)	(0.18)	(0.18)
Class mean first semester grades in Algebra/Russian		−0.00	0.08	0.06		−0.72***	−0.69***	−0.59**
		(0.24)	(0.27)	(0.26)		(0.27)	(0.25)	(0.25)
% girls in class		0.14	0.15	0.08		0.20	0.35	0.31
		(0.46)	(0.46)	(0.47)		(0.42)	(0.43)	(0.42)
% students with >100 books in home		−0.05	−0.07	−0.05		−0.20	−0.17	−0.04
		(0.41)	(0.40)	(0.40)	NO	(0.38)	(0.39)	(0.39)
Teacher experience and category	NO	NO	YES	YES	NO	NO	YES	YES
Constant	0.07	−0.21	−0.46	−0.65	0.08	2.77***	2.71***	2.56***
	(0.10)	(0.95)	(1.28)	(1.31)	(0.12)	(0.96)	(0.90)	(0.89)
Observations	169	169	169	169	173	173	173	173
R-squared	0.05	0.06	0.11	0.14	0.07	0.13	0.18	0.21

Notes: robust standard errors in parentheses.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Russian test but only indirectly—that performance is much more related to "objective" indicators that we would expect to be important in influencing the USE score; namely students' ability and family cultural capital.

5.3. Teachers' views of their relationship with their class and the accuracy of their predictions for their class' test performance

Our estimates of the difference between teachers' prediction of class scores and the actual class scores in the USE examination suggest that mathematics teachers who reported very good relations with their class tended to overestimate significantly their class average USE results compared to teachers' who reported their relationship with their class as good (see Table 5 column 1–4), even when we control for class characteristics, teacher experience and category level, and the proportion of students who reported a very good relationship with their teacher. The effect is relatively large (about 0.40 standard deviations) and robust.

Russian teachers who reported very good relations with their class also tend to overestimate significantly their class average USE results, but once we control for class characteristics—namely, the average grades received by students in the first semester—the estimated coefficient for teachers' view of their relation with the class is not statistically significant (see Table 5 columns 5–8). However, this appears to result from the high correlation between Russian teachers' view of their relationship with the class and students' previous grades in Russian. Thus, the higher the first semester grades in her class, the greater the teacher's overestimate of her students' average performance on the USE Russian test. Regression to the mean plays a role here, but so does teachers' overly rosy opinion of the abilities of students with high grades in Russian (grades given to them by these same teachers) and students in the advanced track. Whether or not this is the direct effect of Russian teachers' view of their relation with their class or the indirect effect of their view of this relationship based on students' higher grades in Russian, teachers who feel very good about their class tend to overestimate their Russian results on the USE.

Teachers who view their relationship with their class as "neutral or bad" tend to underestimate their class' performance on the USE score. Yet for both mathematics and Russian teachers, the actual score minus predicted score difference (teacher accuracy) is not significantly different from the actual minus predicted score difference of teachers who view their relationship with the class as "good."[7]

6. Discussion

We set out to show that teachers may not be accurate in their assessment of students' academic abilities. Our results using a sample in three Russian regions suggest that even under conditions of teachers having extensive information about how well students should do on an important high stakes test of their abilities, teachers who feel particularly good about their relations with their students greatly overestimate their class' actual performance in mathematics.

Such mis-estimation also characterizes Russian teachers' predictions of student performance on the Russian portion of that high stakes test, but in an indirect way. Russian teachers' view of their relationship to the class is apparently more influenced (compared to their math teacher counterparts) by students' previous grades and less by how students in the class view their relationship to their teacher (Table 2). When we control for students' previous grades in Russian, the accuracy of Russian teachers' estimates of student performance is not significantly related to their view of their relation to the class (Table 5). Further, students' performance on language tests correlates more with previous performance and cultural capital (family books in the home) than for mathematics tests (see Table 4). Correspondingly, teachers in language arts may have less conviction that their own teaching can affect students' performance. Therefore, Russian teachers who feel very good about their relationship with the class overestimate class performance on the USE, but their overestimate is also closely identified with how they have evaluated students' previous performance in Russian language.[8]

On the other hand, teachers who have a particularly low opinion of their relationship with their class are much less likely to err in their expectations of class test performance, either in mathematics or Russian. Although such teachers tend to underestimate how well their students will do on the USE, their estimates are not significantly different from those of teachers with average ("good") opinions of their class relations, and both groups of teachers are accurate in estimating their students' performance (teacher expectations do not differ significantly from actual average class performance).

Are teachers' views of their relationship to their class, and, in turn, the accuracy of their predictions of how well their students do on the USE test, influenced by students' social class or gender? We can only approximate these relations on the basis of class means. It appears that the class' gender composition has no significant relation to teachers' view of their class or to their predictions of results or to their likelihood of error in predicting results. Neither does it appear that students' cultural capital influenced the accuracy of math teachers' predictions of their class' USE performance. On the other hand, Russian teachers seem have been influenced by students' cultural capital in their view their relationship with their class and their prediction of their class' USE test score. Yet cultural capital seemed to have no significant direct influence on the accuracy of teachers' predictions of students' USE performance (Table 5).

The issue of whether Russian and mathematics teachers are biased in their predictions of higher cultural capital students' performance on the USE is more complex. We showed that both Russian and mathematics teachers tended to predict higher performance on the USE test for classes with higher cultural capital, even when we control for their class' average grades (Table 4). However, we also showed that students' cultural capital is significantly related to their actual performance on the USE—this in addition to the positive relation to actual score of their previous grades. Thus, teachers' higher predictions of higher cultural capital classes' performance on the USE may not be a source of "bias," but simply reflect a correct assessment of the influence of both first semester grades and cultural capital on how much students know about the subject or how well they will likely do on the test.

The most interesting implications of our results are that teachers who claim to have very good relations with their students are likely to have inaccurate predictions/expectations of how well

[7] The constant term for the estimated difference of actual minus predicted score is not significantly different from zero for both math and Russian teachers when only the teachers' view dummy variables are included in the regression. This implies that teachers who report a "good" relationship with their students tend to accurately estimate their class' average USE test score.

[8] Another possible explanation for this phenomenon is that Russian language study in the 11th grade is largely USE preparation and communications skills development, whereas in mathematics, new (and rather difficult) material is introduced even in the second semester of 11th grade, and these new topics are covered in the USE exam. So predicting USE results based on previous performance (first semester grades, for example) may be more difficult for math teachers.

their students will perform on high stakes assessments of their subject knowledge. That could mean that students with high teacher evaluations will regress to the mean on any given test, or it could mean that teachers who think they have good relations with their class simply overestimate the effectiveness of their teaching or incorrectly assume that good relations with students means that they are learning more. For teachers of Russian, this inaccuracy is interwoven with a likely bias that they have regarding students with higher cultural capital. Teachers of Russian report better relations with classes where students have higher cultural capital, and this, in turn, fuels inaccurate (or biased) estimates of how well such students will do on the USE tests.

Acknowledgement

The authors gratefully acknowledge financial assistance from the Basic Research Program of the National Research University Higher School of Economics.

References

Auwarter, A.E., Aruguete, M.S., 2008. Effects of student gender and socioeconomic status on teacher perceptions. J. Educ. Res. 101 (4), 242–246. doi:http://dx.doi.org/10.3200/JOER.101.4. 243-246.

Bishop, J.H., 1997. The effect of national standards and curriculum-based exams on achievement. Am. Econ. Rev. 87 (2), 260–264.

Bourdieu, P., 1986. The forms of capital. Handbook of Theory and Research for the Sociology of Education. Greenword Press, New York, pp. 241–258.

Bourdieu, P., Passeron, J.C., 1976. Reproduction in Education, Society and Culture. Sage, London; Newbury Park, CA.

Brophy, J.E., 1983. Research on the self-fulfilling prophecy and teacher expectations. J. Educ. Psychol. 75 (5), 631–661. doi:http://dx.doi.org/10.1037/0022-0663.75.5.631.

Carnoy, M., Rothstein, R. (2013). What do international tests really show about U.S. student performance? Retrieved from http://www.epi.org/publication/us-student-performance-testing/.

Cohen, D.K. (Ed.), 1993. Teaching for Understanding: Challenges for Policy and Practice JosseyBass, San Francisco.

Darling Hammond, L., 1996. The right to learn and the advancement of teaching: research, policy, and practice for democratic education. Educ. Res. 25 (6), 5–17.

Dee, T.S., 2005. A teacher like me: does race, ethnicity, or gender matter? Am. Econ. Rev. 95 (2), 158–165.

Dusek, J.B., Joseph, G., 1983. The bases of teacher expectancies: a meta-analysis. J. Educ. Psychol. 75 (3), 327–346. doi:http://dx.doi.org/10.1037/0022-0663.75.3.327.

Ferguson, R.F., 2003. Teachers' perceptions and expectations and the black–white test score gap. Urban Educ. 38 (4), 460–507. doi:http://dx.doi.org/10.1177/0042085903038004006.

Jussim, L., 1991. Grades may reflect more than performance: comment on Wentzel (1989). J. Educ. Psychol. 83 (1), 153–155. doi:http://dx.doi.org/10.1037/0022-0663.83.1.153.

Jussim, L., Harber, K.D., 2005. Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. Pers. Social Psychol. Rev. 9 (2), 131–155. doi:http://dx.doi.org/10.1207/s15327957pspr0902_3.

Kelly, S., Carbonaro, W., 2012. Curriculum tracking and teacher expectations: evidence from discrepant course taking models. Social Psychol. Educ. 15 (3), 271–294.

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., Palumbo, P., 1998. The accuracy and power of sex social class, and ethnic stereotypes: a naturalistic study in person perception. Pers. Social Psychol. Bull. 24 (12), 1304–1318. doi:http://dx.doi.org/10.1177/01461672982412005.

McKown, C., Weinstein, R.S., 2008. Teacher expectations, classroom context, and the achievement gap. J. School Psychol. 46 (3), 235–261. doi:http://dx.doi.org/10.1016/j.jsp.2007.05.001.

Oakes, J., 1985. Keeping Track: How Schools Structure Inequality. Yale University Press, New Haven, CT.

Page, R., 1987. Teachers' perceptions of students: a link between classrooms, school cultures, and the social order. Anthropol. Educ. Q. 18 (2), 77–99. doi:http://dx.doi.org/10.1525/aeq.1987.18.2.04x0667q.

Qing, L., 1999. Teachers' beliefs and gender differences in mathematics: a review. Educ. Res. 41 (1), 63–76.

Ready, D.D., Wright, D.L., 2011. Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities the role of child background and classroom context. Am. Educ. Res. J. 48 (2), 335–360. doi:http://dx.doi.org/10.3102/0002831210374874.

Riegle-Crumb, C., Humphries, M., 2012. Exploring bias in math teachers' perceptions of students' ability by gender and race/ethnicity. Gend. Soc. 26 (2), 290–322. doi:http://dx.doi.org/10.1177/0891243211434614.

Rosenthal, R., 1994. Interpersonal expectancy effects: a 30-year perspective. Curr. Dir. Psychol. Sci. 3 (6), 176–179.

Rosenthal, R., Jacobson, L., 1968. Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development. Holt, Rinehart and Winston, New York.

Rosenthal, R., Rubin, D.B., 1978. Interpersonal expectancy effects: the first 345 studies. Behav. Brain Sci. 1 (03), 377–386.

Rubie-Davies, C., Hattie, J., Hamilton, R., 2006. Expecting the best for students: teacher expectations and academic outcomes. Br. J. Educ. Psychol. 76 (3), 429–444. doi:http://dx.doi.org/10.1348/000709905x53589.

Shepardson, D.P., Pizzini, E.L., 1992. Gender bias in female elementary teachers' perceptions of the scientific ability of students. Sci. Educ. 76 (2), 147–153. doi:http://dx.doi.org/10.1002/sce.3730760204.

Tach, L.M., Farkas, G., 2006. Learning-related behaviors, cognitive skills, and ability grouping when schooling begins. Social Sci. Res. 35 (4), 1048–1079. doi:http://dx.doi.org/10.1016/j.ssresearch.2005.08.001.

Tenenbaum, H.R., Ruck, M.D., 2007. Are teachers' expectations different for racial minority than for European American students? A meta-analysis. J. Educ. Psychol. 99 (2), 253–273. doi:http://dx.doi.org/10.1037/0022-0663.99.2.253.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., Holland, R.W., 2010. The implicit prejudiced attitudes of teachers: relations to teacher expectations and the ethnic achievement gap. Am. Educ. Res. J. 47 (2), 497–527. doi:http://dx.doi.org/10.3102/0002831209353594.