TEACHING PRACTICES AND STUDENT ACHIEVEMENT: EVIDENCE FROM TIMSS

Jan-Christoph Bietenbeck

Master Thesis CEMFI No. 1104

December 2011

CEMFI
Casado del Alisal 5; 28014 Madrid
Tel. (34) 914 290 551. Fax (34) 914 291 056
Internet: www.cemfi.es

This paper is a revised version of the Master's Thesis presented in partial fulfillment of the 2009-2011 Master in Economics and Finance at the Centro de Estudios Monetarios y Financieros (CEMFI). I would like to thank my thesis adviser, David Dorn, for his guidance throughout this project. I am grateful to Manuel Bagues, Jesse Rothstein and participants in the final thesis presentation for their helpful comments. All errors are my own.

Master Thesis CEMFI No. 1104
December 2011

TEACHING PRACTICES AND STUDENT ACHIEVEMENT: EVIDENCE FROM TIMSS

Abstract

Despite a wide consensus among researchers and practitioners that teachers matter to student achievement in schools, the exact determinants of teacher quality remain elusive. This paper follows a recent line of research and examines the impact of different teaching practices on student test scores in the United States. It does so against the background of a series of proposed teaching reforms which call for an increase in the use of "modern" teaching practices at the cost of more "traditional" ones, thus implicitly assuming that the former are better at raising student achievement. Using student survey data from the 2007 wave of the Trends in International Mathematics and Science Study and an estimation strategy which allows me to control for the subject-invariant part of unobserved student ability, I find evidence which points in the opposite direction. While my traditional-teaching measure has a substantial positive effect on student achievement, the estimated impact of my modern-teaching measure is much smaller and statistically insignificant. This result is robust to a series of robustness checks.

Jan-Christoph Bietenbeck
CEMFI
j.bietenbeck@gmail.com

1 Introduction

Teachers matter. This is the consensus from a wide range of studies which examine the impact of teachers on student outcomes. Nevertheless, which teacher attributes in particular make the difference between a successful teacher and an unsuccessful one remains unclear. Variables which are commonly observed in data sets such as teacher education and experience are generally found to have only little impact on student achievement (Hanushek, 1986). This is disquieting not least because these characteristics are typically the main determinants of teacher salary and hiring decisions (Hanushek and Rivkin, 2006). In a renewed attempt to elicit "what makes an effective teacher" (as in Lavy, 2011), a recent line of research therefore shifts the focus from teacher attributes to teaching practices, that is, what teachers actually do in the classroom (Lavy, 2011; Schwerdt and Wuppermann, 2011). The intuition behind this is that differences in instructional methods may be the reason for the large empirically observed variation in teacher quality. If this is the case, straightforward and potentially cost-effective policy changes, such as instructing teachers to teach in a particular way, could help raise student achievement in schools.

In the United States, the last two decades have seen an unprecedented surge in proposals for teaching reform from a variety of sources, including national teacher associations (e.g., National Council of Teachers of Mathematics, 1991) and the National Research Council (1996). Many of these proposals have also been funded by the Department of Education (Zemelman et al., 2005). Given this diversity in authorship, the recommendations made are remarkably congruent. In particular, a common element among almost all of these proposals is the appeal to reduce the reliance on "traditional" teaching practices such as lecture-style teaching and rote memorization, and to increase instead the use of more "modern" teaching methods including cooperative group work among students and teaching based on student questioning (*ibid.*). The implicit assumption behind these proposed teaching reforms, which are jointly referred to as the standards movement in teaching practices, is thus that modern teaching practices are better than traditional ones at raising student achievement - an assumption which has not been tested empirically so far.[1]

This paper attempts to fill this gap in the literature and to thereby contribute to the still sparse evidence on the link between teaching practices and student achievement.

[1] Schwerdt and Wuppermann (2011) study whether teachers who emphasize lecture-style teaching as opposed to problem solving are associated with higher student achievement. These two practices can however not be considered representative of traditional and modern teaching as defined by the standards movement for reasons that will become clear below.

Using student survey data from the latest wave of the Trends in International Mathematics and Science Study (TIMSS), I construct two aggregate teaching-practice measures, one for traditional teaching and one for modern teaching. I then relate these measures to student test scores from standardized tests in math and science using an identification strategy which allows me to control for the subject-invariant part of unobserved student ability. My results suggest that while there is a substantial positive impact of traditional teaching on student achievement, the impact of modern teaching is much smaller and statistically insignificant. While I cannot reject the hypothesis that the effect size of both measures is equal, my results do not support the hypothesis that modern teaching is better at raising student achievement than traditional teaching either. This casts doubt on the usefulness of the recommendations made by the standards movement.

A series of robustness checks, including a redefinition of my teaching-practice measures and the use of highly flexible econometric specifications, confirm the validity of my headline results. When I examine the effect of the two measures on math and science achievement separately, I find that the pattern of a larger positive impact of traditional teaching holds for both subjects. I also analyze whether the treatment effects differ for various subgroups of my sample. I find that the effects are roughly equal across boys and girls. In contrast, there is some evidence pointing towards a more favorable effect of modern teaching for immigrant students. The remainder of the paper is structured as follows: Section 2 briefly reviews the most relevant literature. Section 3 presents the data. The empirical strategy is described in Section 4. Section 5 presents the headline results. Robustness checks are discussed in Section 6. Section 7 concludes.

2 Related Literature

A wealth of research tries to link measurable teacher characteristics to student outcomes using observational data. The typical approach to this problem is to set up an education production function in which student achievement as measured by some form of standardized test is related to teacher experience, education and certification. Evidence from the associated regressions points towards a positive effect of teacher certification on test scores (Clotfelter et al., 2010; Dee and Cohodes, 2008). However, teacher experience and education - the variables most frequently used to inform hiring and salary decisions - are generally found to have no significant effect on student achievement (Hanushek, 1971, 1986; Hanushek and Rivkin, 2006). Evidence from recent

studies which examine the impact of teacher gender is also inconclusive (Dee, 2007; Holmlund and Sund, 2008). An alternative approach that has become popular during the last decade therefore avoids the focus on particular teacher characteristics and attempts to identify a generic measure of teacher quality instead. Studies which adopt this approach typically exploit longitudinal data sets in which teachers face different groups of students over time. In an equation of student achievement gain, a value-added measure of teacher quality is then calculated as a teacher fixed effect.[2] Researchers conclude that there is substantial variation in teacher quality and that its impact on student achievement is large. However, consistently with the evidence from the education production function approach, teacher experience and education are found to explain only very little of the variation in estimated teacher quality (Aaronson et al., 2007; Rivkin et al., 2005).

The elusiveness of measurable determinants of teacher quality and the availability of new and richer data has prompted researchers to shift the attention from teacher characteristics to what teachers do in the classroom very recently. Two studies from this emergent literature are particularly closely related to this paper. First, Schwerdt and Wuppermann (2011) use data from the TIMSS 2003 wave for the United States to contrast the effect of lecture-style teaching with that of solving problems in class on standardized test scores. The authors find that teachers who spent relatively more time on lecture-style teaching are associated with higher student achievement. Second, Lavy (2011) uses student survey data from Israel to examine the effect of five aggregate teaching practices on standardized test scores. He finds that two of these practices, "instilment of knowledge" and "instilment of applicative, analytical and critical skills," which he likens to "traditional teaching" and "modern teaching," respectively, are positively related to student achievement. The author concludes that traditional and modern teaching approaches do not necessarily crowd out each other as is commonly thought, but that both may coexist in the education production function.

In keeping with this recent line of research, this paper provides additional evidence on the link between teaching practices and student achievement. Similar to Lavy (2011), I use student survey data from the United States to construct two aggregate teaching-practice measures for traditional and modern teaching. I then relate these measures to standardized test scores in math and physics. While Schwerdt and Wuppermann (2011) also study the effect of teaching practices in the United States, this paper extends beyond their work in two ways. First, while their variable of inter-

[2] Value-added models of teacher quality have come under harsh criticism recently (Rothstein, 2010). I discuss these objections in further detail in Section 4 of this paper.

est is the relative intensity of two teaching practices, my data contains a large amount of teaching practices, which allows me to adopt a broader definition of traditional and modern teaching. Second, since this definition is in line with that of the standards movement, I am able to directly evaluate its policy recommendations.

3 Data

TIMSS is an international assessment of the math and science knowledge of fourth- and eighth-grade students. It was first carried out in 1995 by the International Association for the Evaluation of Educational Achievement (IEA) and has been repeated every four years thereafter. For reasons that will become clear below, this study focuses on the 2007 sample of eighth-grade students for the United States, which consists of 7377 individuals in 235 schools. TIMSS 2007 sampled students in a two-stage clustered sampling design, in which schools were selected in the first stage, and two math classes were randomly sampled within each of these schools in the second stage.[3] Within each sampled class, in principle all students participated in the assessment. In practice, however, the number of sampled students may be smaller than the actual class size because of student nonparticipation (Williams et al., 2009). To account for this complex sampling design, sampling weights need to be applied and a jackknife resampling technique be employed to calculate standard errors correctly in statistical analysis.

Participating students were administered standardized tests in math and science. The tests consisted of both multiple-choice questions and constructed-response items, the latter of which required students to generate and write their own answers. In practice, the use of an incomplete-booklet design implied that each individual student only completed a random subset of items from a larger pool of questions. Each student's test scores for the overall test were then imputed from her responses, and are made available in the data in the form of five imputed values (also called *plausible values*). In addition to the overall math and science scores, TIMSS reports test scores for the four subsections of the science test - chemistry, physics, biology, and earth science - for each student. I use information on science course content from the teacher questionnaire to select each student's corresponding test score to be used in my analysis (for instance, for a student whose teacher reports to have taught a physics class I select her physics test score).[4] I also standardize test scores in each subject, with mean zero and standard

[3] If there were no more than two eighth-grade math classes in a given school, all of these classes were selected with certainty.

[4] The idea behind this is that a physics teacher's teaching practices should not influence her students'

deviation equal to one. This lets me interpret the estimated coefficients as fractions of a standard deviation of the test score distribution.

Besides assessing students' achievement in math and science, TIMSS collects detailed background information from students, their teachers and their school principals via questionnaires. In particular, the 2007 wave of TIMSS asked students to rate on a four-point scale how often they do a range of different activities in their math and science classes.[5] The answers are coded such that a value of 1 corresponds to "never", 2 to "some lessons", 3 to "about half of the lessons", and 4 to "every or almost every lesson." To compare these activities to the teaching practices referred to in the standards movement, I make use of a listing included in Zemelman et al. (2005). This listing, which is the result of a survey of the standards movement literature, categorizes teaching practices either as recommended "to be decreased" or as recommended "to be increased" separately for math and science. Table A1 gives an overview of some of the items in the listing.

I am able to unambiguously match six math activities and seven science activities from the TIMSS student questionnaire to the teaching practices in Table A1. I group activities categorized as "to be decreased" under the heading "traditional teaching," and those categorized as "to be increased" under the heading "modern teaching." The matched class activities and their grouping are displayed in Table A2.[6] Note that some of the activities are common to math and science - for instance, the math activity "We memorize formulas and procedures" corresponds to the science activity "We memorize science facts and principles" - while others do not have an obvious counterpart in the other subject. In the main part of my analysis, I use all of the matched activities for each subject to create my two teaching-practice measures of interest. Later, I redefine the measures to include only matched activities which are common to both subjects and repeat my analysis using these measures as a robustness check.

For each subject and heading, I calculate the mean of each student's answers across

performance in, say, the biology part of the science test. I should note that the variable from which I draw the information on science course content contains a large number of missing values. In order not to reduce my sample size too much, I select the overall science score for students whose teachers did not provide the necessary information.

[5]Note that in contrast to Lavy (2011), where students answer to which share of their teachers a particular teaching practice applies, students here respond separately for each subject (and thus separately for each teacher, as will become clear below).

[6]Note that from the information in Table A1, Schwerdt and Wuppermann's (2011) "lecture-style teaching" activity can be unambiguously matched to the traditional-teaching category in Table A2. However, it is not clear where their comparison activity "problem solving" belongs in the standards movement classification: for instance, solving routine problems is considered a practice "to be decreased", while tackling complex problems which require new solution paths is "to be increased."

activities, and then compute the class-level mean of these means for each student while excluding the student's own (mean) answer.[7] The resulting measures of traditional teaching and modern teaching share a common scale, which naturally ranges from 1 to 4. Unfortunately, the categorical nature of students' answers implies that the two measures do not stand in a mechanical trade-off to each other: scoring one point higher on the traditional-teaching measure does not necessarily imply that the modern-teaching measure decreases by one point. In fact, correlation plots reveal that the two measures are positively correlated both at the student and at the class level. Nevertheless, these measures let me reasonably address my question of interest, namely whether modern teaching is indeed better than traditional teaching at raising student achievement as the standards movement supposes.

Out of the initial 7377 individuals in my sample, 295 either could not be linked to their teachers or have more than one teacher in math or science. I exclude these students from my analysis. Moreover, I restrict my sample to classes with at least three students and drop individuals with missing information on the two teaching-practice measures of interest (that is, I drop students of whom no classmate answered the questions used to construct the teaching-practice measures). This leaves me with 6843 students in 234 schools. For ease of exposition in the following discussion, I will refer to this reduced sample as the full sample. Missing values for control variables are a common concern in survey data, and the TIMSS data are no exception. I therefore specify parsimonious sets of control variables at the student-, teacher- (i.e. class-), and school level and delete observations with missing information on any of these variables. The resulting estimation sample consists of 4642 eighth-grade students with 271 math teachers and 303 science teachers in 182 schools.

Table 1 shows descriptive statistics for my traditional-teaching and modern-teaching measures, teacher controls, and class controls separately for math and science classes and for both the full sample of 6843 students and the reduced estimation sample of 4642 students. Mean differences between subjects are reported in the respective final column of each sample's panel. Comparing the statistics for the full sample with those of the estimation sample reveals no great differences between the two samples. This makes me confident that the reduced estimation sample is still representative of the target population of eighth-grade students in the United States. The following discussion is based on the figures from the estimation sample as these are of the most interest to me.

[7] I also experimented with including the student's own answer in the class-level mean and with not aggregating answers at the class level at all. In both cases, my results did not change much.

Examining first the two teaching-practice measures, I find that the means of traditional teaching in math and science are exactly equal at 3.07 points. In contrast, modern teaching in math is 0.15 points lower than in science at 2.66 points, a difference which is statistically significant. The teacher controls used in my regressions are drawn from the TIMSS teacher questionnaires and comprise a relatively standard set of variables: a gender dummy, four categorical age dummies, three dummies for teaching experience, and a dummy for teachers who majored in the field they teach during their studies. Note that I decided to group teacher experience into three categories even though it is reported as a continuous variable in the data set. This categorization reflects the commonly observed fact that any gains in student achievement associated with teaching experience take place in the first five years, with the largest part of these gains occurring in the first year (e.g., Clotfelter et al., 2010; Harris and Sass, 2011). The two class controls included in my regressions are the number of minutes per week that the subject is taught (drawn from the teacher questionnaire) and the number of students observed in the class. As the last column of Table 1 reports, the share of female teachers, the number of minutes per week taught, and the number of students are significantly higher in math classes than in science classes.

Table 2 displays descriptive statistics for student controls and school controls separately for the full sample and the estimation sample. As with the teacher controls in Table 1, a comparison of the means and standard deviations reveals no large differences between the two samples. The student controls used in my regressions are drawn from the TIMSS student questionnaire. They comprise a gender dummy, the student's age, two dummies for black and hispanic students, a dummy for students born outside the United States, and a dummy for students who report that English is not the primary language spoken at home. Moreover, I include in the controls the number of books at home as a proxy for parental background since the parental education variable contained many missing values and its inclusion would have drastically reduced my sample size. Cross-correlations of the two variables reveal that as expected, higher parental education is correlated with a higher reported number of books at home.[8] I also include in my regressions the following school controls, which I draw from the TIMSS school questionnaire that was given to participating schools' principals: three dummies of parental involvement, a dummy indicating whether more than fifty percent of students at the school were eligible for free lunch, and the total student enrollment in grade eight.

[8] I nevertheless repeated my regressions including parental education instead of the number of books at home. This did not change my results much.

4 Empirical Strategy

The challenge in using observational data to identify the causal effect of a particular teaching practice on student achievement is to deal with the potential nonrandom assignment of students and teachers to classrooms. If students with high unobserved ability are systematically paired with teachers using a particular teaching practice, for example, then the estimated coefficient on this practice will be biased upward. Studies linking other teacher characteristics to student outcomes have typically addressed this issue by using panel data on students, where the fact that individuals are observed for several consecutive periods allows one to introduce student fixed effects which control for time-invariant unobservables at the student level. The matched-pairs nature of the TIMSS data - students are observed twice, once in math and once in science - lets me use a related identification strategy: between-subject differencing. As Rothstein (2010) points out in the context of panel data, the use of student fixed effects does not resolve the sorting problem when *time-varying* unobservable determinants of student achievement are correlated with classroom assignment. The equivalent concern with between-subject differencing is that *student ability may be subject-specific* (Clotfelter et al.,2010). I discuss this issue in more detail below.

My analysis parts from a relatively standard education production function, which relates student i's test scores in subject $j\varepsilon\{m,s\}$ in school k, y_{ijk}, to the teaching-practice measures of interest, TP_{ij}, student traits, X_i, teacher and class characteristics, T_j, and school characteristics, S_k:

$$y_{ijk} = \alpha_j + TP'_{ij}\beta_{1j} + X'_i\beta_{2j} + T'_j\beta_{3j} + S'_k\beta_{4j} + \varepsilon_{ijk}. \tag{1}$$

The error term ε_{ijk} contains the unobservable determinants of student i's test score. In particular, it can be written as the sum of student unobservables, η_{ij}, teacher and class unobservables, τ_j, school unobservables, θ_{kj}, and an idiosyncratic part, ν_{ijk}:

$$\varepsilon_{ijk} = \eta_{ij} + \tau_j + \theta_{jk} + \nu_{ijk}. \tag{2}$$

Note that equation (2) allows student- and school unobservable determinants to vary across subjects.

If I estimated equation (1) by ordinary least squares, any correlation between the unobservable determinants in the error term and the teaching-practice measures would cause a bias in the estimate of β_1. I can eliminate some of these potential sources of

bias by assuming that unobserved student and school traits are equal across subjects, that is, $\eta_{im} = \eta_{is}$ and $\theta_{im} = \theta_{is}$. Differencing between subjects then leads to:

$$\begin{aligned} y_{imk} - y_{isk} \quad = \Delta y_i = \quad & (\alpha_m - \alpha_s) + TP'_{im}\beta_{1m} - TP'_{is}\beta_{1s} + X'_i(\beta_{2m} - \beta_{2s}) \\ & + T'_{im}\beta_{3m} - T'_{is}\beta_{3s} + S'_i(\beta_{4m} - \beta_{4s}) \\ & + (\tau_m - \tau_s) + (\nu_{imk} - \nu_{isk}). \end{aligned} \tag{3}$$

That is, differencing eliminates unobserved student and school traits and thus controls for between- and within-school sorting of students. Note that identification of β_1 in equation (3) relies on the variation of the teaching-practice measures across subjects for each student. This motivates the focus on eighth-grade students in my analysis: fourth-grade students in the United States typically have a single teacher for all subjects, which implies that the necessary between-subject variation in teaching practices does not exist for them.

A typical assumption made at this point (e.g., Dee, 2007; Schwerdt and Wuppermann, 2011) is that all observables influence student achievement equally across subjects, that is, $\beta_{.m} = \beta_{.s}$. This implies that observable student and school traits drop out of equation (3), and that the rest of the terms may be summarized to yield:

$$\Delta y_i = (\alpha_m - \alpha_s) + (TP'_{im} - TP'_{is})\beta_1 + (T'_{im} - T'_{is})\beta_3 + (\tau_m - \tau_s) + (\nu_{imk} - \nu_{isk}). \tag{4}$$

While I also estimate equation (4), my headline specification allows the coefficients on student and school observables to differ between subjects. That is, my headline specification reads

$$\begin{aligned} \Delta y_i = \quad & (\alpha_m - \alpha_s) + (TP'_{im} - TP'_{is})\beta_1 + X'_i\beta_2 \\ & + (T'_{im} - T'_{is})\beta_3 + S'_i\beta_4 + (\tau_m - \tau_s) + (\nu_{imk} - \nu_{isk}). \end{aligned} \tag{5}$$

Bias in this specification may arise from any of three sources. First, if student unobservables are subject-specific (e.g., subject-specific ability), $\eta_{im} - \eta_{is}$ remains in the error term, and if correlated with the difference in the teaching-practice measures will cause a bias in the estimated coefficient of β_1. As an example, if students who are unobservably more able in say, math, are systematically assigned to math teachers that emphasize traditional teaching practices, the estimated coefficient on the traditional-teaching measure will be biased upward. Related to this concern is an implicit assumption in the between-subject identification strategy: since my data do not allow me to

control for prior achievement in the education production function, I *de facto* assume that students' initial knowledge in each subject is negligible. While I cannot address these concerns definitely with the TIMSS data - that is, I cannot test the assumption that $E(\eta_{im} - \eta_{is})(TP'_{im} - TP'_{is}) = 0$ - I should emphasize again that any overall ability effect on achievement is captured by my identification strategy.

Second, a similar argument can be made for unobservable teacher characteristics in case $E(\tau_m - \tau_s)(TP'_{im} - TP'_{is}) \neq 0$. That is, even after differencing between subjects, teacher unobservables $\tau_m - \tau_s$ remain in the error term. If correlated with both student achievement and the teaching-practice measures, this again will generate a bias in the estimate of β_1. As an example, it might be the case that more motivated teachers sort into modern teaching practices. If teacher motivation is at the same time positively related to student achievement, this will lead to an upward bias in the estimated coefficient on the modern-teaching measure. Again, my data do not allow me to test the assumption that $E(\tau_m - \tau_s)(TP'_{im} - TP'_{is}) = 0$. Finally, a third concern is that teachers may adjust their teaching practices according to the students they face. This very plausible idea casts doubt on the source of variation in the teaching-practice measures. As a consequence, I refrain from interpreting my estimates as causal effects. Rather, I advocate an interpretation of β_1 as a measure unlikely to be driven by between- or within-school sorting, but which may partly be determined by teacher sorting into particular teaching practices based on unobservable teacher characteristics.

5 Results

Table 3 shows the results of the between-subject estimation. All regressions in this and the successive tables include five dummies to control for teacher-reported science course content, one each for integrated or general science, chemistry, physics, biology, and earth science.[9] I start out with a very basic specification using only the teaching-practice measures as explanatory variables in column 1, and successively add more control variables in the following columns.[10] In particular, the results in column 2 cor-

[9] Given the large amount of missing information in the corresponding teacher-questionnaire variable, I incorporated nonresponse into the "integrated or general science" category. While this practice may be criticized, the fact that my results hardly change when I do not include any science course-content dummies implies that my conclusions do not depend on this step. The data also include a variable on math course content, which however suffers from even more severe teacher nonresponse. I therefore refrain from including any dummies for math course content in my regressions.

[10] I also experimented with including each of the measures separately in the regressions. The resulting coefficient estimates were very similar to the ones presented in the main text.

respond to the model in equation (4), and my headline results in column 4 correspond to the model in equation (5). The dependent variable in each of the regressions is the within-student difference between her standardized math and science test score.

The estimated coefficients on both teaching-practice measures are positive across all specifications. It is interesting to see that the coefficient on traditional teaching hardly changes with the addition of controls: its effect size fluctuates around ten percent of a standard deviation of the test score distribution, and it is significant at the 5 percent level in all four regressions. In contrast, the estimated coefficient on modern teaching decreases from 0.033 in column 1 to 0.023 in my headline specification and it is not statistically significant at any common significance level in any specification. Consistently with the previous literature, most of the control variables at the teacher level are found to have small and statistically insignificant effects on student achievement. The only exception is that teachers who majored in the field they teach positively influence student achievement in that subject. This finding is in line with that of Dee and Cohodes (2008), but not with that of Goldhaber and Brewer (1997) who find no significant effect of a subject-specific major. Moving to the class controls, the effect of total teaching time has the expected sign - more teaching time is associated with higher student achievement - but its estimated coefficient is small and statistically insignificant. Finally, the number of students does not have any statistically significant impact on student achievement in my headline specification.

The results discussed in the previous paragraph give no support to the implicit claim by the standards movement that modern teaching is better than traditional teaching at raising student achievement. In fact, taken at face value the estimates suggest that the opposite is true. A natural question is therefore whether the estimated effect sizes of the two teaching-practice measures are statistically distinguishable. I test this as a linear hypothesis using a Wald test; the last row of Table 3 presents the corresponding p value for each specification. It turns out that across all specifications, the p value is too large to reject the null of equal effects at any common significance level. This means that I cannot conclude that traditional teaching is better than modern teaching. At the same time, however, my findings imply that modern teaching is not better than traditional teaching either, a result which casts doubt on the usefulness of the standard movement's recommendations.

Two more points are worth noting with respect to the results in Table 3. First, a natural concern is that the observed effects may be driven by some sort of student or teacher sorting. As discussed in Section 4, my data do not allow me to address this

concern definitively. However, the fact that the coefficient estimates do not change much with the inclusion of additional controls may be seen as evidence that sorting is not the underlying mechanism here: if one believes that the selection on unobservables reflects this selection on observables, one may conclude that my results are unlikely to be confounded by unobserved student, teacher, or school traits. Second, my results are in line with the previous findings of the literature. Like Lavy (2011), I find that more traditional teaching is associated with higher student achievement. This result is also consistent with that by Schwerdt and Wuppermann (2011), who proxy for traditional teaching by lecture-style teaching. While I do not find a statistically significant positive effect of modern teaching on test scores as Lavy (2011) does, my results do not point in the opposite direction either.

6 Robustness Checks and Heterogeneous Effects

Starting from my headline specification in equation (5), in this section I implement a total of six robustness checks which address potential reservations about the validity of the results presented in Table 3. Later, I also examine whether there is heterogeneity in the treatment effects across different subgroups of my sample. In the top panel of Table 4, I present estimation results from regressions which use alternatively-defined versions of my two teaching-practice measures.[11] First, I redefine the measures so as to include only those activities from the student questionnaire which are common to math and science teaching (that is, I only include activities marked $^{+}$ in Table A2). I thereby address a potential concern about the comparability of the measures across subjects. As the results in the left column show, my findings are robust to this redefinition: the estimated coefficient on the traditional-teaching measure hardly changes compared to the one in column 4 of Table 3, while the coefficient on the modern-teaching measure decreases to basically zero.

Second, in the construction of my teaching-practice measures and in the interpretation of my results I implicitly assumed that student answers are measured on a linear scale. However, arguably the difference in teaching time needed to employ a teaching practice "about half the lessons" instead of "some lessons" is smaller than the difference in teaching time needed to employ it "every or almost every lesson" instead of "about half the lessons." To address this concern, I recode the student answers in the

[11] For the sake of conciseness, Table 4 only reports the estimated coefficients on the teaching-practice measures and their standard errors.

data, assigning a value of 5 instead of 4 to the "every or almost every lesson" category. I then construct my teaching-practice measures as described in Section 3 and include them in a regression of equation (5). The results, which are displayed in the right column of the top panel of Table 4, show that both the estimated coefficient on traditional teaching and the one on modern teaching are now smaller than in my headline results. Nevertheless, the combination of a relatively large and statistically significant coefficient on traditional teaching and a small and statistically not significant coefficient on modern teaching carries over to these results. I therefore conclude that my findings do not depend on the exact scaling of students' answers.

In the middle panel of Table 4, I implement two sample restrictions. First, a teacher's choice of her teaching practices may be constrained by the teaching time she has available. If this is the case, large differences in teaching time between math and science classes may drive my findings. To test this hypothesis, I limit my sample to students with a between-subject difference in teaching time of at most two hours per week. As the results in the left column show, repeating my regression for this restricted sample does not change my results much, implying that the hypothesis does not hold. Second, another concern might be that peer effects, whose importance is widely acknowledged in the education literature, account for my results. To address this issue, I restrict my sample to classes with the same student composition in math and science. If peers influence student achievement equally in both subjects - and there is no reason to believe that this is not the case - the between-subject differencing takes care of this effect.[12] Re-estimating equation (5) for this restricted sample, I find that the results, which are reported in the right column of the middle panel of Table 4, differ only slightly from the findings from my headline specification. I therefore conclude that peer effects are not the driving mechanism of my results.

In the bottom panel of Table 4, I relax the assumptions on the coefficients in my model in two ways. First, the left column presents estimated coefficients for the teaching-practice measures from a specification in which I allow β_3 to vary across subjects (that is, I allow teacher controls to influence achievement in math and science differently). Again, the findings do not differ much from my headline results in Table 3. Second, and more interestingly, I estimate a specification in which I let β_1 vary across subjects, thus allowing the teaching-practice measures to have different impacts on math and science achievement. The results of this regression are reported in the right column of the bottom panel of Table 4. Focusing first on the estimated coefficients for math, I find that

[12] More formally, if I introduced a peer fixed effect in my education production function, this effect would drop out in the between-subject differenced equation.

the coefficient on traditional math teaching is approximately fifty percent higher than the main effect of traditional teaching in my headline specification. To obtain a better idea of this effect size, the coefficient implies that moving students from the minimum exposure (2) to the maximum exposure (4) of traditional math teaching observed in the data is associated with a thirty percent of a standard deviation increase in their math test score ((4-2)*0.151). In contrast, and consistently with my headline story, modern math teaching seems to have almost no effect on student achievement.[13]

Turning to the estimated coefficients for science, note that since all science variables enter with a minus sign in equation (5), a negative coefficient implies a positive relationship between the teaching practice and student achievement in science. The estimated coefficient on traditional science teaching is only about half the size of the main effect of traditional teaching in my headline results, and only a third of the size of the corresponding math effect. Moreover, the effect is not statistically significant at any common significance level. In contrast, the coefficient on modern science teaching is about fifty percent higher than the corresponding main effect at 0.036. I therefore conclude that the pattern from my headline results - traditional teaching has a larger estimated effect than modern teaching on student achievement - also exists for math and science teaching separately. However, this effect seems to be driven mainly by the large positive impact of traditional math teaching on students' math test scores.

Finally, I examine whether the treatment effects are heterogeneous across different subgroups of my sample. For this purpose, I first split my sample by students' gender and run the regression in equation (5) separately for boys and girls. The results, which are shown in the left panel of Table 5, display no large differences in the effect sizes between the two groups. In a second step, I split my sample by students' origin, distinguishing between those born in United States and those born abroad. Given the small number of foreign-born students in my sample (406 out of 4642 students), the regression results for the USA-born students, which are shown in the right panel of Table 5, simply reflect my headline results. In contrast, the results for the foreign-born students point in an interesting direction: while the estimated coefficient on traditional teaching is basically zero for this group, the coefficient on modern teaching is about three times as large as in my headline results. This suggests that my headline story of a more favorable effect of traditional teaching is reversed for foreign-born students (although

[13]Note that in a Wald test of equality of the effect sizes of modern and traditional math teaching, I can now reject the null of equal effects at the 10 percent level. As Table 4 shows, however, similar tests for the other robustness checks in this section did not lead to any further rejections of the null at any common significance level.

I should note that none of the coefficients is statistically significant due to the small sample size). In search for an explanation for this effect, I re-estimated equation (5) for students who reported that English is not the main language spoken at home. The results, which I do not report here, were in line with those of my headline specification, which suggests that language difficulties cannot explain the more favorable effect of modern teaching for foreign-born students.

7 Conclusion

Recent proposals for teaching reform in the United States advocate the decrease of "traditional" and the increase of "modern" teaching practices, thereby implicitly assuming that the latter are better at raising student achievement in schools. This paper examines empirically whether this assumption holds and presents evidence which points in the opposite direction: aggregate measures of teaching practices constructed from student surveys indicate that while traditional teaching has a large positive effect on student achievement, the same is not true for modern teaching. This result is robust to a serious of robustness checks, and is unlikely to be driven by between- or within-school sorting of students. However, the empirical strategy used in this paper does not allow me to control for unobserved teacher effects. As a consequence, I refrain from interpreting my results as causal and do not formulate any policy recommendations. My analysis further suggests that for foreign-born students, modern teaching may have a more favorable effect than for USA-born students. This result is unlikely to be due to language difficulties by the former group, and more research is needed to disentangle the underlying reason for this finding. Finally, this paper adds to the still sparse evidence on the importance of teaching methods for student outcomes, and its results point towards teaching practices as a potentially important determinant of teacher quality.

References

Aaronson, D., L. Barrow and W. Sander (2007): "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics*, 25, 95-135.

Clotfelter, C. T., H. F. Ladd and J. L. Vigdor (2010): "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects," *Journal of Human Resources*, 45, 655-681.

Dee, T. S. (2007): "Teachers and the Gender Gaps in Student Achievement," *Journal of Human Resources*, 42, 528-554.

Dee, T. S. and S. R. Cohodes (2008): "Out-of-Field Teachers and Student Achievement: Evidence from Matched-Pairs Comparisons," *Public Finance Review*, 36, 7-32.

Goldhaber, D. D. and D. J. Brewer (1997): "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity," *Journal of Human Resources*, 32, 505-523.

Hanushek, E. A. (1971): "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data," *American Economic Review*, 61, 280-288.

Hanushek, E. A. (1986): "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24, 1141–1177.

Hanushek, E. A. and S. G. Rivkin (2006): "Teacher quality." In *Handbook of the Economics of Education (Vol. 2)*, ed. by E. A. Hanushek and F. Welch, North-Holland, Amsterdam.

Harris, D. N. and T. R. Sass (2011): "Teacher Training, Teacher Quality and Student Achievement," *Journal of Public Economics*, 95, 798-812.

Holmlund, H. and K. Sund (2008): "Is the Gender Gap in School Performance Affected by the Sex of the Teacher?," *Labour Economics*, 15, 37-53.

Lavy, V. (2011). "What Makes an Effective Teacher? Quasi-Experimental Evidence," NBER Working Paper 16885.

National Council of Teachers of Mathematics (1991): *Professional Standards for Teaching Mathematics*. Report, Reston (VA).

National Research Council (1996): *National Science Education Standards*. Report, Washington, D.C.

Rivkin, S. G., Hanushek, E. A. and J. F. Kain (2005): "Teachers, Schools, and Academic Achievement," *Econometrica*, 73, 417-458.

Rothstein, J. (2010): "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics*, 125, 175–214.

Schwerdt, G. and A. C. Wuppermann (2011): "Is Traditional Teaching Really All That Bad? A Within-Student Between-Subject Approach," *Economics of Education Review*, 30, 365-379.

Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., Smith, C. and P. Stearns (2009): "TIMSS 2007 U.S. Technical Report and User Guide," National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Zemelman, S., Daniels, H. and A. Hyde (2005): *Best Practice. Today's Standards for Teaching and Learning in America's Schools (3rd edition)*. Heinemann, Portsmouth (NH).

Table A1
Categorization of Teaching Practices

	Math	Science
Practices to be decreased	Rote memorization of rules, formulas, and procedures. Teaching by telling.	Memorizing detailed vocabulary, definitions, and explanations without thorough connection to broader ideas. Instruction based mainly on lecture and information giving. Dependence on textbooks and lockstep patterns of instruction.
Practices to be increased	Cooperative group work. Justifying answers and solution processes. Connecting mathematics to other subjects in the real world. Word problems with a variety of structures and solution paths.	Collaborative small-group work. Students' reflection to realize concepts and processes learned. Application, either to social issues or further scientific questions. Observation activity, often designed by students, aimed at real discovery, employing a wide range of process skills.

Note: The items presented here are extracted from Zemelman et al. (2005), who summarize the recommendations made by the standards movement in teaching practices over the past two decades.

Table A2
Matched Student Questionnaire Items

	Math	Science
Traditional teaching	We memorize formulas and procedures.†	We memorize science facts and principles.†
	We listen to the teacher give a lecture-style presentation.†	We listen to the teacher give a lecture-style presentation.†
		We read our science textbooks and other resource materials.
Modern teaching	We work together in small groups.†	We work in small groups on an experiment or investigation.†
	We relate what we are learning in mathematics to our daily lives.†	We relate what we are learning in science to our daily lives.†
	We explain our answers.†	We give explanations about what we are studying.†
	We decide on our own procedures for solving complex problems.	We design or plan an experiment or investigation.
		We make observations and describe what we see.

Note: Students responded to the question, "How often do you do these things in your mathematics lesson (in your science lesson)?" Answers are coded on a four-point scale, with 1 corresponding to "never", 2 to "some lessons", 3 to "about half the lessons", and 4 to "every or almost every lesson." Items marked † are considered to be common among math and science and are included in the redefined teaching-practice measures which are used in the robustness checks.

Table 1
Descriptive Statistics: Teaching-Practice Measures, Teacher Controls, and Class Controls

	Full Sample (N=6843)					Estimation Sample (N=4642)				
	Math		Science		Difference	Math		Science		Difference
	Mean	SD	Mean	SD		Mean	SD	Mean	SD	
Teaching-practice measures										
Traditional teaching	3.07	0.31	3.07	0.33	0.00	3.08	0.31	3.06	0.33	0.01
Modern teaching	2.68	0.30	2.81	0.37	-0.13***	2.66	0.30	2.81	0.37	-0.15***
Teacher controls										
Female teacher	0.69	0.46	0.58	0.49	0.11**	0.70	0.46	0.60	0.49	0.10*
Teacher younger than 30	0.20	0.40	0.15	0.36	0.05	0.21	0.41	0.16	0.36	0.06
Teacher aged 30 – 39	0.29	0.45	0.30	0.46	-0.01	0.26	0.44	0.29	0.45	-0.03
Teacher aged 40 – 49	0.26	0.44	0.25	0.43	0.01	0.27	0.45	0.25	0.43	0.03
Teacher older than 49	0.25	0.43	0.30	0.46	-0.05	0.26	0.44	0.31	0.46	-0.05
Teaching experience <1 year	0.06	0.24	0.06	0.24	0.00	0.06	0.23	0.06	0.24	0.00
Teaching experience 1 – 5 years	0.20	0.40	0.23	0.42	-0.03	0.20	0.40	0.22	0.41	-0.02
Teaching experience >5 years	0.74	0.44	0.71	0.45	0.03	0.74	0.44	0.72	0.45	0.02
Teacher majored in field taught	0.46	0.50	0.39	0.49	0.08*	0.50	0.50	0.45	0.50	0.05
Class controls										
Number of students in class	15.7	4.5	13.1	5.2	2.6***	15.6	4.4	12.9	5.0	2.7***
Teaching time (min/week)	247.4	78.7	230.5	62.8	16.9***	246.6	79.2	231.3	63.6	15.3***

Note: Each parameter estimate presented in the "Difference" columns is obtained from a separate regression. */**/*** denotes significance at the 10/5/1 percent level.

Table 2
Descriptive Statistics: Student Controls and School Controls

	Full Sample[a]		Estimation Sample[a]	
	Mean	SD	Mean	SD
Student controls				
Female	0.51	0.50	0.51	0.50
Age	14.30	0.48	14.30	0.48
Black	0.12	0.33	0.12	0.33
Hispanic	0.23	0.42	0.21	0.41
Foreign-born	0.10	0.29	0.08	0.27
English not main language at home	0.09	0.29	0.08	0.27
Number books at home: 0 – 10	0.17	0.38	0.16	0.37
Number books at home: 11 – 25	0.21	0.40	0.21	0.41
Number books at home: 26 – 100	0.28	0.45	0.28	0.45
Number books at home: 101 – 200	0.17	0.37	0.17	0.38
Number books at home: >200	0.17	0.38	0.17	0.38
School controls				
Parental involvement: low	0.33	0.47	0.31	0.46
Parental involvement: medium	0.43	0.50	0.46	0.50
Parental involvement: high	0.23	0.42	0.22	0.42
Share free-lunch eligible >50%	0.41	0.49	0.38	0.49
Total enrollment grade eight	257.9	181.4	247.8	169.9

[a] Number of observations is 6843 in the full sample and 4642 in the estimation sample.

Table 3
Estimation Results: Between-Subject Differencing

	(1)	(2)	(3)	(4)
Traditional teaching	0.107***	0.096**	0.101**	0.096**
	(0.044)	(0.043)	(0.042)	(0.043)
Modern teaching	0.033	0.030	0.010	0.023
	(0.037)	(0.037)	(0.036)	(0.037)
Female teacher		0.002	0.008	0.009
		(0.022)	(0.021)	(0.021)
Teacher younger than 30		0.042	0.040	0.031
		(0.036)	(0.036)	(0.035)
Teacher aged 30 – 39		-0.016	-0.014	-0.019
		(0.023)	(0.022)	(0.023)
Teacher older than 49		-0.010	-0.016	-0.019
		(0.024)	(0.023)	(0.022)
Teaching experience <1 year		-0.007	-0.010	0.003
		(0.044)	(0.043)	(0.043)
Teaching experience 1 – 5 years		-0.027	-0.033	-0.031
		(0.029)	(0.029)	(0.029)
Teacher majored in field taught		0.035*	0.035*	0.036*
		(0.020)	(0.020)	(0.019)
Teaching time (min/week) x 10^{-3}		0.098	0.116	0.117
		(0.135)	(0.134)	(0.137)
Number of students in class		0.007**	0.006*	0.004
		(0.003)	(0.003)	(0.003)
Student controls			✔	✔
School controls				✔
Observations	4642	4642	4642	4642
Average R^2	0.005	0.010	0.034	0.038
H_0: Traditional = Modern (p value)	0.260	0.298	0.142	0.247

Note: The dependent variable is the difference between math and science test scores. Each regression is run five times (once for each plausible value); the second but last row presents the average R^2 from these regressions. All regressions include dummies for science course content. Variables included in student and school controls are listed in Table 2. See text for the definition of the traditional-teaching and modern-teaching measures. Standard errors in parentheses are calculated using the appropriate jackknife procedure and allow for clustering at the school level. */**/*** denote significance at the 10/5/1 percent level.

Table 4
Robustness Checks

Redefinition of teaching-practice measures

	Only common elements	"Almost always" equals 5
Traditional teaching	0.101 (0.049)*	0.068 (0.030)**
Modern teaching	0.004 (0.033)	0.007 (0.027)
Observations	4642	4642
Average R^2	0.037	0.036
H_0: Traditional = Modern (p value)	0.129	0.178

Sample restrictions

	Δ teaching time <2 hours	Same peers
Traditional teaching	0.098 (0.047)*	0.110 (0.047)*
Modern teaching	0.019 (0.042)	0.014 (0.050)
Observations	3747	2775
Average R^2	0.039	0.039
H_0: Traditional = Modern (p value)	0.242	0.207

Heterogeneous effects

	β_3 varies across subjects	β_1 varies across subjects
Traditional teaching	0.099 (0.048)*	
Modern teaching	0.013 (0.037)	
Traditional teaching math		0.151 (0.046)***
Traditional teaching science		-0.053 (0.064)
Modern teaching math		0.006 (0.054)
Modern teaching science		-0.036 (0.040)
Observations	4642	4642
Average R^2	0.039	0.038
H_0: Traditional = Modern (p value)	0.211	
H_0: Trad. math = modern math (p value)		0.054
H_0: Trad. science = modern science (p value)		0.84

Note: all regressions are variations of the specification in column 4 of Table 3 – see the corresponding note to that table for further information. For details of the restrictions imposed in these regressions, see text.

Table 5
Heterogeneous Effects

	Gender		Country of birth	
	Girl	Boy	USA	Abroad
Traditional teaching	0.092 (0.055)*	0.104 (0.062)	0.106 (0.047)**	0.001 (0.122)
Modern teaching	0.011 (0.045)	0.027 (0.063)	0.014 (0.037)	0.072 (0.099)
Observations	2372	2270	4236	406
Average R^2	0.030	0.026	0.038	0.073
H_0: Traditional = Modern (p value)	0.308	0.440	0.160	0.686

Note: all regressions are variations of the specification in column 4 of Table 3 – see the corresponding note to that table for further information.

MASTER'S THESIS CEMFI

0801 *Paula Inés Papp*: "Bank lending in developing countries: The effects of foreign banks".

0802 *Liliana Bara*: "Money demand and adoption of financial technologies: An analysis with household data".

0803 *J. David Fernández* Fernández: "Elección de cartera de los hogares españoles: El papel de la vivienda y los costes de participación".

0804 *Máximo Ferrando Ortí*: "Expropriation risk and corporate debt pricing: the case of leveraged buyouts".

0805 *Roberto Ramos*: "Do IMF Programmes stabilize the Economy?".

0806 *Francisco Javier Montenegro*: "Distorsiones de Basilea II en un contexto multifactorial".

0807 *Clara Ruiz Prada*: "Do we really want to know? Private incentives and the social value of information".

0808 *Jose Antonio Espin*: "The "bird in the hand" is not a fallacy: A model of dividends based on hidden savings".

0901 *Víctor Capdevila Cascante*: "On the relationship between risk and expected return in the Spanish stock market".

0902 *Lola Morales*: "Mean-variance efficiency tests with conditioning information: A comparison".

0903 *Cristina Soria Ruiz-Ogarrio*: "La elasticidad micro y macro de la oferta laboral familiar: Evidencia para España".

0904 *Carla Zambrano Barbery*: "Determinants for out-migration of foreign-born in Spain".

0905 *Álvaro de Santos Moreno*: "Stock lending, short selling and market returns: The Spanish market".

0906 *Olivia Peraita*: "Assessing the impact of macroeconomic cycles on losses of CDO tranches".

0907 *Iván A. Kataryniuk Di Costanzo:* "A behavioral explanation for the IPO puzzles".

1001 *Oriol Carreras:* "Banks in a dynamic general equilibrium model".

1002 *Santiago Pereda-Fernández:* "Quantile regression discontinuity: Estimating the effect of class size on scholastic achievement".

1003 *Ruxandra Ciupagea:* "Competition and "blinders": A duopoly model of information provision".

1004 *Rebeca Anguren:* "Credit cycles: Evidence based on a non-linear model for developed countries".

1005 *Alba Diz:* "The dynamics of body fat and wages".

1101 *Daniela Scidá:* "The dynamics of trust: Adjustment in individual trust levels to changes in social environment".

1102 *Catalina Campillo:* "Female labor force participation and mortgage debt".

1103 *Florina Raluca Silaghi:* "Immigration and peer effects: Evidence from primary education in Spain".

1104 *Jan-Christoph Bietenbeck:* "Teaching practices and student achievement: Evidence from TIMSS".