What Makes an Effective Teacher? Quasi-Experimental Evidence*

July 2011

Victor Lavy

Department of Economics, Hebrew University of Jerusalem

University of Warwick, NBER and CEPR

Abstract

This paper measures empirically the relationship between classroom teaching practices and student achievements. Based on primary- and middle-school data from Israel, I find very strong evidence that two important elements of teaching practices cause student achievements to improve. In particular, classroom teaching that emphasizes the instilment of knowledge and comprehension, often termed "traditional"-style teaching, has a very strong and positive effect on test scores, particularly among girls and pupils of low socioeconomic background. Second, the use of classroom techniques that endow pupils with analytical and critical skills ("modern" teaching) has a very large positive payoff, evidenced in improvement of test scores across subgroups differentiated by gender and socioeconomic background. However, the effect of each of these two teaching-practices are different at different treatment intensity, the first has its highest effect at low to medium levels of treatment, while the second has its largest impact at high levels of treatment. I also find that transparency, fairness, and proper feedback in teachers' conduct with their students improve academic performance, especially among boys. However, I find no evidence of an effect of a second element of modern teaching, instilment of the capacity for individual study. Apart from identifying "what works" in the classroom, these findings yield two insights for the debate about the merit of "traditional" versus "modern" approaches to teaching, which are often discussed as rival classroom pedagogical approaches. First, one approach does not necessarily crowd out the other; both may coexist in the classroom production function of knowledge. Second, it is best to target the two teaching practices differentially to students of different genders and abilities. The effect of the effective teaching practices estimated is very large, especially in comparison with that of other potential interventions such as reducing class size or increasing school hours of instruction.

JEL No. I2, I21, J24

---

* I thank Michael Beenstock, Daniel Paserman, Steve Pischke, Jonah Rockoff, Yona Rubinstein and seminar participants in numerous institutions and conferences for most useful discussions and suggestions. Special thanks go to Daniel Schuchalter and Agnia Galesnik for their excellent research assistance.

1. Introduction

While teacher quality may be important, it is driven by characteristics that are difficult or impossible to measure. This is often the conclusion of many past and recent studies that failed to produce consistent evidence linking pupils' achievement to observable teacher characteristics (Hanushek, 1986). As an alternative, researchers have tried recently to separate student achievements into a series of "fixed effects," assigning importance to teachers, schools, pupils, and so on. For example, Rockoff (2004), Rivkin et al. (2005), Kane et al. (2008b), and Aaronson et al. (2007) demonstrate substantial and persistent variations in achievement growth among students assigned to different teachers. An even more recent strand of research tries anew to identify specific characteristics of teachers that affect pupils' achievements. In contrast to older studies that examined mostly the effect of teachers' demographic and educational characteristics , these new studies (e.g., Kane, Rockoff and Staiger, 2008b, Rockoff et al., forthcoming, Rockoff and Staiger, forthcoming) focus on characteristics such as cognitive ability, content knowledge, personality traits, and personal beliefs regarding self-efficacy.

In this paper, I shift the attention from teachers' personal characteristics and attributes to what they do in the classroom in an attempt to identify the most effective teaching practices. In particular, I measure teaching practices on the basis of data obtained from surveys among primary- and middle-school students in Israel in 2002–2005. Students are asked about the intensity that their teachers engage in specific activities. The idea that students can characterize what their teachers are doing in the classroom raises a number of measurement and interpretation issues which I discuss in the data and methodology sections. However, other researchers and organizations recently endorsed and implemented this approach in other settings. For example, The Bill &Melinda Gates Foundation project "Measures of Effective Teaching," launched in fall 2009 also uses a survey among students to measure perceptions of the classroom instructional environment (Bill and Melinda Gates Report, 2010) and their effect on students' value added.[1]

I use the conceptual categorization of teachers' pedagogical practices as developed in the educational-psychology literature (Bloom, 1956) to summarize information based on 29 dimensions of pedagogy in five aggregated measures of teaching practices. These measures are (1) instilment of knowledge and enhancing comprehension; (2) instilment of applicative, analytical, and critical skills; (3) instilment of the capacity for individual study; (4) transparency, fairness, and feedback, and (5) individual treatment of students.[2] I use three different ways of aggregating the students' responses and

[1] The Gates Foundation Tripod survey, like the Israeli surveys, asks students if they agree or disagree with a variety of statements about their teachers' classroom practices. Many of its questions resemble those in the Israeli survey.

[2] The Tripod questions in The Bill &Melinda Gates Foundation project are summarized in eight principal components. One of them, "Consolidate," strongly resembles Bloom's definition of "instilment of knowledge and enhancement of comprehension." A second component, "Challenge," largely overlaps Bloom's "instilment of

then I examine which of these characteristics cause students' test-score outcomes to improve. Even though individually only a few of the 29 teaching characteristics have statistically significant effects on student outcomes, the first two factors of these teaching practices clearly have a large and statistically significant effect on students' test scores. I also find significant effects for some of the other factors by allowing for heterogeneity in treatment by students' gender and socioeconomic status.

This paper makes its main contribution by identifying multiple ways of measuring effective teaching for pupils' learning. The evidence that I present here, however, also provides insights of relevance for the ongoing policy debate about the relative merit of "traditional" versus "modern" methods of classroom teaching. In some countries, recent policy action has replaced traditional teaching methods with modern ones or vice versa. In the U.S., for example, the National Standards recommend modern teaching practices that engage students in self- and group-learning activities (National Council of Teachers of Mathematics, 1991, National Research Council, 1996).[3] In England, conversely, Michael Gove, the Secretary of Education, announced shortly after taking up his post in the summer of 2010 a reform that would reintroduce traditional teaching and learning in schools.[4] In the summer of 2008, the Israel Ministry of Education unveiled an opposite reform.[5] The empirical evidence that I present in this paper addresses these policy initiatives directly, since "instilment of knowledge and enhancement of comprehension" includes a variety of classroom practices that are viewed as the core of traditional teaching, requiring students to acquire knowledge through drill, practice, and memorization (Salomon,

---

applicative, analytical and critical skills." The other three principal components that I use do not overlap perfectly with the other components in the Gates Foundation report although, as I noted above, many other similar individual items appear in both. See Bill and Melinda Gates Report (2010) for further detail.

[3] Zemelman, Daniels, and Hyde (1993 and 2005) provide a normative typology of teaching practices for schools in the U.S. Traditional practices that should be decreased, they say, include rote practice, rote memorization of rules and formulas, single answers and single methods of finding answers, the use of drill worksheets, repetitive written practice, teaching by telling, teaching computation out of context, stressing memorization, testing for grades only, and being the dispenser of knowledge. Modern teaching practices that should be put to greater use are manipulative materials, cooperative group work, discussion of mathematics, questioning and making conjectures, justification of thinking, writing about mathematics, a problem-solving approach to instruction, content integration, use of calculators and computers, facilitating learning, and assessing learning as an integral part of instruction.

[4] For details of the recent reform, see, for example, the *Daily Mail* of November 25, 2010. Michael Gove's prescription for the improvement of state schools (expressed while he was still the shadow Secretary of Education) focuses on return to learning based on memorization and comprehension, e.g., reciting multiplication tables, learning to conjugate verbs, and memorizing important dates and figures in national history. Gove also claimed, "It is often the poorest children who suffer most from trendy teaching. When synthetic phonics was abandoned as the way of teaching children to read because it was too authoritarian, children from book-rich backgrounds survived but those who were already book poor fell behind." (See the *Daily Telegraph*, October 20, 2007, Rachel Sylvester and Alice Thomson interview with England's shadow Education Secretary, and, more recently, Michael Gove article in the *Times*, March 17, 2010.)

[5] 'Pedagogical Horizon,' a 2008 circular from the Ministry Director General, advised that as of September 1, 2008, teaching at the post-primary level would change from being based on memorization, repetition, and practice to emphasis on development of deep understanding and acquisition of learning and thinking skills. The facilitation of pedagogical reform entailed other changes, including in curriculum and standards, teacher training, and student evaluation.

Perkins, and Theroux, 2001). The "instilment of applicative, analytical, and critical skills" measure captures the main elements of modern teaching style, focusing on the instilment of learning skills and creative thinking (Resnick, 1987).

I use in the empirical analysis panel data on pupils in Israel who are in fifth grade in 2002 (primary school) and in eighth grade (middle school) in 2005. These students were tested in four subjects (English, Hebrew, mathematics, and science) in both grades as part of a national testing program. I base the identification on within-pupil analysis (using pupil fixed effects) together with primary- and middle-school fixed effects that net out any confounding factors among schools or individuals. Therefore, variations in teaching practices within schools/grades and across classes are natural variations in teaching styles of teachers. I use this heterogeneity to estimate the effect of interest and demonstrate below that it does not correlate with any of the pupil or class characteristics that may affect potential outcomes. In practice, the within-pupil estimation eliminates only some of the selection or sorting of pupils into primary and middle schools as some of the estimates remain unchanged while others do change when primary- and middle-school fixed effects are added to the estimated equations. Additional evidence that supports the causal interpretation of the findings presented in this paper is the institutional rules that forbid school choice and tracking by ability in primary and middle public schools in Israel and that classes are actually formed randomly. The findings about heterogeneity of the effect of treatment, by type of teaching practices and by types of students, specifically by gender and ability, support the claim about the causal nature of our finding because boys and girls, for example, are sorted similarly across schools and among classes within schools. If the results are produced by sorting, they should be similar for males and females, but they are not. This heterogeneity in treatment effects constitute also an evidence against the possibility that it is overall teacher effectiveness and not specific teaching practices that is causing students' value added.

As mentioned, the evidence presented in this paper suggests considerable pupil heterogeneity by gender and socioeconomic background in the effect of both teaching practices. Thus, while instilment of knowledge and enhancing comprehension has a generally positive effect, it is much larger for girls and for pupils from less-educated families; conversely, the effect of the instilment of applicative, analytical, and critical skills is positive for both genders and both levels (low and high) of student socioeconomic status. The coexistence of positive and heterogeneous effects of these seemingly contradictory teaching practices has important policy implications for the potential improvement of pupils' knowledge by the targeting of teaching methods. Another interesting outcome of the study is that the three other teaching practices studied have no systematic relation with the improvement of pupils' test scores.

The effect size of the two statistically significant teaching practices is very large relative to other educational interventions such as reducing class size or improving teacher training. For example, if

pupils are moved from the minimum to the maximum exposure observed in the sample of each of these two teaching-practice measures, the average test score in each subject increases by one standard deviation of the test-score distribution. The "traditional teaching" practice accounts for 0.60 of this change; the "modern teaching" practice contributes 0.40. A more realistic effect-size simulation, based on improving these two teaching practices from the mean to the best observed in the data, leads to an improvement of about 0.50 of a standard deviation in the test score in each subject. These large effect sizes of teaching practices and the consistency of the findings reported in this paper are important as many countries search for ways to promote "teacher quality" or "teacher effectiveness." This ardent interest, partly occasioned by anxiety in some countries about students' poor performance in international tests and comparisons such as TIMMS, PISA and PIRLS, is not being addressed by much reliable evidence. The findings reported in this paper are a step toward meeting this demand.

The paper proceeds as follows: Section 2 presents brief review of the most relevant literature, Section 3 describes the data, and Section 4 explains the empirical methodology. The results, robustness checks, and heterogeneity in the treatment effects of teaching practices are presented and discussed in Section 5. The last section concludes and suggests policy implications.

2. Related Literature

A lengthy list of studies represents the efforts of researchers to use non-experimental data to estimate teacher effects on pupils' learning outcomes (e.g., Hanushek, 1971; Murnane and Phillips, 1981; Rockoff, 2004; Hanushek, Rivkin, and Kain, 2004; Jacob and Lefgren, 2005; Aaronson, Barrow, and Sander, 2007; Kane, Rockoff, and Staiger, forthcoming; Gordon, Kane, and Staiger, 2006, Cantrell et al 2008). Several additional studies used random assignment to estimate the variation in teacher effects. In this kind of analysis, Nye, Konstantopoulous, and Hedges (2004) revisited the results of the STAR experiment in Tennessee. After accounting for the effect of different classroom-size groupings, their estimate of the variance in teacher effects was well within the range typically reported in the non-experimental literature. Chetty et al. (2010), based on the STAR experiment data as well, followed the project participants to adulthood and have shown that students who had a more experienced teacher in kindergarten have higher earnings.

Some researchers (e.g., Clotfelter et al., 2006, 2007) found that teachers with stronger academic backgrounds produce larger performance gains for their pupils; others did not find this relationship (e.g., Harris and Sass, 2006, on graduate coursework and Kane et al., 2008a, on college selectivity). A small number of studies (e.g., Clotfelter et al., 2006, 2007; and Goldhaber, 2007) found a link between teachers' scores on certification examinations and their effectiveness; Harris and Sass (2006) found no such relation.

Researchers acknowledge the possibility that non-random assignment of students to teachers may distort measures of teacher effectiveness. Some teachers are given better students who would achieve well in many different classrooms. Some researchers question whether a teacher's specific contribution can be accurately estimated at all, given the possibility that students are assigned to teachers on the basis of unmeasured characteristics not captured by test scores and demographics (Rothstein, 2010). Other researchers, while recognizing the potential of bias, are more optimistic (Koedel and Betts (2007). One recent study (Kane, Rockoff and Staiger, 2008b) compared experimental (i.e., classes randomly assigned to teachers) and non-experimental estimates of teachers' effects on student achievement growth for a small sample of teachers in Los Angeles. In that sample, the non-experimental or observational measures predicted the experimental measures with little bias—as long as the observational models controlled for each student's prior achievements.

In several studies, the effect of teachers in one grade fades out as students progress to higher grades (McCaffrey et al, , 2004; Kane, Rockoff and Staiger, 2008a; Jacob, Lefgren, and Sims, 2008; Rothstein, 2010). Hypotheses for the fadeout range from artifacts of empirical strategy to the heterogeneity of teacher quality within schools to the relevance of skills gained in one year for skills tested the next year (Kane, Rockoff and Staiger, 2008b).

A few recent studies found a relationship between a teacher's measured effect on student achievements and overall subjective administrator ratings (Jacob and Lefgren, 2005; Rockoff and Speroni, 2010; Rockoff, Jacob, Kane and Staiger, forthcoming). These studies, however, do not identify the criteria or behaviors that the school principals use to make their judgments. Two papers that are close in motivation and measurement of treatment to this study is Tyler et al., (2010) and Kane et al., (2011). The first paper describes the early and the second the final results from the same study where teachers' evaluation and their teaching practices are related to student achievement. The study finds that measures of teaching effectiveness are substantively related to student achievement growth and that some observed teaching practices predict achievement more than other practices.

The Learning about Teaching (Bill and Melinda Gates foundation, December 2010) report about initial findings from Measures of Effective Teaching Project suggests that a teacher's past track record of value-added is among the strongest predictors of their students' achievement gains in other classes and academic years. The teachers who lead students to achievement gains on one year or in one class tend to do so in other years and other classes. The report also suggests that student perceptions of a given teacher's strengths and weaknesses are consistent across the different groups of students they teach and that student perceptions in one class are related to the achievement gains in other classes taught by the same teacher.

3. Data

The empirical analysis uses two samples, one of fifth-grade primary-school students and one of eighth-grade middle-school students. Both samples were culled from Jewish secular schools to the exclusion of others, because this part of the Israeli school system places Grades 5 and 8 in different schools—primary and middle, respectively. The distinction is important because it requires students to change schools in the middle and because the secular school system does not allow school choice at either the primary or the middle level, except in a few cities where charter schools that allow opting out of neighborhood schools. I address this issue again in the Methodology section below.

I drew the two samples from the Growth and Effectiveness Measures for Schools (GEMS)—Meizav in Hebrew—datasets for 2002 and 2005. GEMS, composed of a series of tests and questionnaires administered by the Evaluation and Measurement Division of the Ministry of Education,[6] is administered at the midterm of each school year to a representative 1-in-2 sample of all primary and middle schools in Israel, so that each school participates in GEMS once every two years.

The GEMS student data include test scores of fifth- and eighth-graders in mathematics, science, Hebrew, and English, as well as the responses of fifth- through ninth-grade students to questionnaires. In principle, all students except those in special-education classes are tested and required to complete the questionnaire. The proportion of students tested is above 90% and the rate of questionnaire completion is roughly 91%. The raw test scores use a 1–100 scale that we transform into z-scores to facilitate interpretation of the results. The tests in all four subjects measure facts, analytical skills and also critical thinking, both at low and high level. These properties of the achievement tests are important because we may expect traditional teaching to matter more for learning about facts while we would expect modern practices to be more important for critical thinking.

The GEMS student questionnaire addresses various aspects of the school and learning environment. The section I used, focusing on teaching style and practices, includes 29 items that are listed in the Appendix. These items ask students to rate on a six-point scale to what proportion of their teachers the statement is appropriate. A score of 1 indicates 'none of the teachers', 2 indicates 'very few teachers', 3 indicates 'few teachers', 4 indicates 'part of the teachers', 5 indicates 'large part of the teachers', and 6 indicates 'almost all teachers'. The items refer to the current year's teachers so it is not a retrospective report but on the other hand it has the limitation of ignoring a student's history of teachers. Perhaps we can view this limitation as an advantage here because students, particularly those in primary grades, are less likely to be particularly retrospective. Another point to emphasize is that these survey questions

[6] The GEMS assessments are not administered for school accountability purposes; only aggregate results at the district level are published. For more information on GEMS, see the Evaluation and Measurement Division website (in Hebrew): http://cms.education.gov.il/educationcms/units/rama/odotrama/odot.htm.

provide information about the average teaching practices of all teachers that teach a particular class and not about a particular teacher. We should note this distinction when interpreting the evidence presented later sections. Although we probably care also about the practices of the teacher in the subject of interest, rather than the proportion of a student's teachers who exhibit a certain practice, the latter measure is much less likely to be endogenous. The average teaching styles of teachers of a given classroom is less likely to be a function of student abilities, class size and so on. Nevertheless, I describe in the next section the identification strategy where I control for student's ability by a pupil's fixed effect and I provide evidence that classes are formed randomly out of the students in a given grade. Both of these elements limit very much the possibility that the variation in teaching practices that I use for estimation reflects some endogenous placement of certain teaching practices.

I group the items under five categories that describe teachers' pedagogical practices in the classroom: (1) instilment of knowledge and enhancement of comprehension (seven items); (2) instilment of applicative, analytical and critical skills (nine items); (3) instilment of capacity for individual study (three items); (4) transparency, fairness, and feedback (three items); and (5) individual treatment of students (seven items). These categories of teachers' pedagogical practices correspond to common and accepted terminology in the educational-psychology literature, dating back to Bloom (1956), on major categories in the taxonomy of educational objectives ("Categories in the Cognitive Domain"). Each category comes with detailed list of outcomes and a list of outcome-illustrating verbs. Knowledge, for example, is defined as the remembering (recalling) of appropriate previously learned information; the associated verbs include define, describe, enumerate, and others that are listed in Appendix B.[7] Bloom (1956) defines comprehension as grasping (understanding) the meaning of informational materials; some of the associated verbs in this category are classify, convert, describe, discuss, and explain. Bloom uses both concepts, "knowledge" and "comprehension," to define his first teaching practice, "instilment of knowledge and enhancement of comprehension." In relying on Bloom's logic and categorization, I avoid some arbitrariness in grouping the items in different categories even though some people could disagree with Bloom as to the appropriate placement of certain items. However, it is important to note here that the results reported in this paper are stable with respect moving items that may seem to some as less definitive in terms of the category to which they 'belong'. For example, it can be argued that item 3 in T1 ("The teachers commend students who know the material well") or item 7 in T1 ("I understand the teachers' scholastic requirements well") could legitimately be placed in the transparency, fairness and feedback domain. Changing the locations of these items does not affect the basic results reported below.

[7] Retrieved from (http://faculty.washington.edu/krumme/guides/bloom.html).

Such robustness of the estimated effect of the various categories is also an indication that the number of items included in each category is not affecting the reliability of the composite.

In this paper, I focus on the first four teaching-practice measures and do not report evidence on the fifth measure ("individual treatment of students") because it reflects the level of the students in class and its estimated effect is prone to reverse causality. I do, however, include this measure in all the estimated regressions that I report in the paper even though the estimated effect of the other four teaching-practice measures does not change when the fifth measure is omitted from the estimated equations—partly because the estimates of this measure are almost always small and not significantly different from zero.

I aggregate the student's responses to a class level measure in three different ways. First I treat the teaching practices variables as cardinal treatment and simply average the students' categorical responses to each question, and then use the mean of all questions that form each of the teaching practices as the treatment variables. Since these measures are based on categorical responses, there is no obvious justification for treating the numerical values assigned to the categorical responses as meaningful. Therefore, as an alternative treatment measure I use a dummy variable that indicates whether the class mean of a given teaching practice is above a certain percentile in the sample distribution of that teaching practice. I use five alternative cutoff points: the median, and the 60th, 70th, 80th, and 90 percentile. The third approach of measuring the teaching practices is different from the first two: here I use the proportion of students that give a given or higher categorical response in each of the items that form a teaching practice. I use three thresholds for computing this measure: 4 ('part of the teachers' or more), 5 ('large part of the teachers' or more), and 6 ('almost all teachers'). This is an ordinal measure and it will be useful the compare the results obtained when using it to those I obtain when using the first cardinal measure.

I linked the student questionnaire data and the 2002 and 2005 test scores data to student administrative records collected by the Israel Ministry of Education which include student demographic background characteristics. Using the linked datasets, I built one data set for primary schools and another for middle schools. The primary-school file for 2002 includes data from 415 primary schools with test scores and student questionnaires. The means of students' characteristics in this sample are presented in column 1 of Table 1. The panel sample includes students from 359 primary schools and their mean characteristics are presented in column 2 of Table 1. Restricting the panel sample to primary schools with at least five students leaves 122 such schools and their students' mean characteristics are presented in column 3 of Table 1. The mean characteristics of pupils included in the panel dataset (in column 2 or 3) are not very different from those of the full sample though some of these differences are significant. But note that the differences in father and mother year of schooling as well as the gender composition are very small and they are not statistically different from zero. Therefore, we can conclude that the panel

sample is representative of the full sample in terms of the background characteristics that are most important as determinants of cognitive outcomes.

The middle school file for 2005 includes data from 176 schools with eighth grade students' questionnaires and test scores. The panel sample with at least five students in each school includes 192 schools, 122 primary schools and 70 middle schools.

4. Empirical Strategy

The structure of GEMS makes it possible to track a sample of students from primary schools (fifth grade in 2002) to middle schools (eighth grade in 2005).[8] I used this feature to construct a longitudinal dataset at the student level to examine how changes in teaching practices (styles and methods) induce changes in pupils' test scores. We know from prior work (e.g., Rivkin et al., 2005) that there exists dramatic cross-grade variation in measures of teacher quality and I use in this paper such variation for estimating the effect of teaching practices. However, it is important to note that that this change is due to the compulsory transition of students from primary to middle school. Also important is the fact that Israel does not allow school choice at the primary and middle levels; pupils are assigned to their neighborhood primary school and middle school, the latter often having a catchment area that includes several primary schools.

Since the estimated regression includes a student fixed effect and a school fixed effect, the identification is based on contrasting the change in exposure to the various teaching practices during grades five and eight among students who followed the same transition path from primary to middle school. More formally, I assume that the cognitive achievements of pupils in grades five and eight are determined by the following equation:

$$(1) \qquad y_{ics} = \alpha_i + \beta_s + S'_{cs}\lambda_2 + \sum_\tau \theta_\tau (TeachingPractice)_{\tau sc} + \varepsilon_{ics} + \in_{cs}$$

where *i* denotes individuals, *c* denotes class (within a grade), *and s* denotes schools. Since the school indicator is perfectly correlated with the grade indicator (fifth or eighth grade), there is no need to add a grade effect in Equation (1). y_{ics} is an achievement measure for student *i* in class *c* and school *s*; α_i is a pupil effect, β_s is a school effect, S'_{cs} is a vector of characteristics of class *c* in school *s*; it includes a set of variables for average characteristics of students in the class (mother's and father's years of schooling, number of siblings, immigration status, and five groups of ethnic origin), characteristics of class learning environment and climate (such as levels of classroom noise, violence, lack of discipline and class size). $(TeachingPractice)_{\tau sc}$ is a vector of four teaching practices $(\tau = 1...4)$ in class *c* and school *s*. The

[8] I did not link datasets from consecutive years because almost all localities were sampled once every two years.

error term in the equations is composed of a school-specific random element $\in_{cs}$ that allows for any type of correlation within observations of the same school across classes and an individual random element ε_{ics}. The coefficient of interest is θ, which captures the effects of the different teaching practices. For the purpose of comparison, I will also present OLS estimate of regressions that do not include pupil's fixed effects but include instead individual characteristic as controls:

$$(2) \qquad y_{ics} = \alpha + \beta_s + x'_{ics}\lambda_1 + S'_{cs}\lambda_2 + \sum_\tau \theta_\tau (TeachingPractice)_{\tau sc} + \varepsilon_{ics} + \in_{cs}$$

Where x'_{ics} is a vector of student's covariates that includes mother's and father's years of schooling, number of siblings, immigration status, and ethnic origin, and indicators for missing values in these covariates.

To estimate Equations (1) and (2), I need to observe students while they are in fifth and eighth grade. For the estimates in Equation (1) to have a causal interpretation, however, the unobserved determinant of achievement must be uncorrelated with the treatment variable. The inclusion of school fixed effects and pupil fixed effects controls for the most obvious potential confounding factor—the endogenous sorting of students across schools. However, there may be unobserved within-school and across-class factors that also correlate with changes in teachers' teaching practices. If some classroom characteristics are not controlled, the estimated effects of interest will be biased. Random assignment of students and teachers to classrooms solves this problem by breaking the link between teaching practices and extraneous effects on the class such as unobserved peer quality. True random-assignment variation is rare in an education context and unavailable in many countries. However, students in Israel's primary and middle schools are rarely grouped into classes on the basis of ability or family background; in fact, such practices are forbidden by law. Therefore, classes in primary schools with multiple classrooms at the same grade level are typically formed on a more-or-less random basis; classes in middle schools are formed in a way that creates social integration by mixing students from different socioeconomic backgrounds.[9] Since all classes within a grade are of equal average ability, teachers are assigned to classes more-or-less randomly and the possibility of better teachers avoiding assignment to lower-performing classes is irrelevant, as is the possibility for "teacher-shopping" by parents. I note here also that the lack of tracking in primary and middle schools in Israel rule out as well the possibility that class composition changes across subjects. Therefore, the students in a given class rank the same teachers.

[9] A 1968 education reform established a three-tier structure of schooling in Israel: primary (grades 1–6), middle (7–9), and high (10–12). The reform established neighborhood school zoning as the basis of primary enrollment and integration, sometimes with busing, of students out of their neighborhoods in middle school. Tracking and sorting of students in primary- and middle-school classes were outlawed and the law is strictly enforced.

The foregoing implies that $(TeachingPractice)_{\tau sc}$ will be uncorrelated with class-level shocks $\in_{cs}$ conditional on a set of school fixed effects, pupil fixed effects, and class-mean characteristics. Thus, the basic identifying assumption in this study is that the systematic components of teaching practices in school arise only at the school level and not at the class level. A necessary condition for the within-school estimation to work is, of course, that there is sufficient variance in teaching practices within a school, as is the case in our data.

The identification strategy I use in this paper is most closely related to that of Ammermueller and Pischke (2009), who use a school-fixed-effects framework to estimate peer effect based on within-school and across-class variation in peer ability. They demonstrate that conditioned on a school fixed effect, class composition within a grade is random. However, I also include a pupil fixed effect in the regressions, which accounts for any selection based on pupil specific attributes. This addition is very important in the context of this paper because it is possible that the teaching practices vary based on students' ability. The student fixed effects that I include in the regression are therefore appropriate controls that rule out a bias due to potential endogeneity of teaching practices. Including also in the regressions the class level characteristics ($S_{cs}^{'}$) is also useful in this regard.

Evidence of the Validity of the Identification Strategy

The key identifying assumption I make in this paper postulates that, conditional on pupil fixed effects, changes in teaching practices within a school are uncorrelated with changes in unobserved factors that may affect students' outcomes. I assess here, from different angles, the plausibility of this assumption. I first discuss the assignment of students both between and within schools and present evidence that sheds light on the question of whether classes are formed (more-or-less) randomly and whether different classrooms systematically get different resources. Even if the variation in teaching practices within a school resembles a random process, however, these variations may be correlated with additional class-to-class changes that may affect student outcomes. To assess this possibility, we check whether changes in teaching practices within a school are associated with changes in student background characteristics such as parental education, family size, ethnicity, and student's immigration status.

Students in Israel attend primary school from initial enrollment to grade 6 and middle school from grades 7 to 9. Generally speaking, primary-school assignment depends on place of residence. Each middle-school catchment area includes several primary schools in order to achieve social integration by blending pupils from different socioeconomic backgrounds. Parents can affect choice of school in certain ways, e.g., by choosing to live near the school of their choice. The school administration is responsible for assigning students to classes within schools. Extra resources are allocated to schools that have a high

share of disadvantaged or recent-immigrant students but class size cannot exceed 40 students in all schools. An important regulation from the Ministry of Education forbids grouping of students by ability in primary and middle school.[10] Even when parents fund additional weekly instruction hours, these resources cannot be used for the formation of study groups by ability (tracking) or any other criterion.[11] A similar regulation from the Ministry applies to middle schools and requires heterogeneous classes. For example, a circular from the Director General outlining the responsibilities of a middle-school principal relative to the responsibilities and authority of a secondary-school principal states explicitly that it is the responsibility of the former to create heteroegenous classes and that ability tracking is allowed only after ninth grade.[12] These institutional rules are supported by evidence from PIRLS 2003, based on the item in the school questionnaire that asks whether the school forms classes on the basis of ability. The fraction of students in schools that report some ability grouping at the class level is close to zero and it does not vary by gender. I obtained similar evidence from TIMSS 1999 and PIRLS 2003, which included a similar question about the extent to which classes are formed based on students' ability. Jakubowski (2009), using PIRLS 2003 data to study the effect of tracking, included Israel among countries that do not track students by ability in their primary- and middle-school systems.

Having obtained this institutional evidence of random formation of classes within schools, I used the sample of all primary and middle schools to test whether the data I use in this study also support this claim. In particular, I checked class assignment to see whether it correlates systematically with students' characteristics. For this purpose, I performed a series of Pearson Chi-Square (χ^2) tests for eight characteristics: gender, father's years of schooling, mother's years of schooling, number of siblings, and three ethnic origin indicators. If a school forms its classes randomly, any particular characteristic of a student should be statistically independent of his or her class assignment. In father's schooling, for example, the Pearson χ^2 test asks whether there are more pupils with high father's schooling in a particular class than is consistent with independence, given the size of the school's enrollment. Ammermuler and Piscke (2009) describe and apply this test in their study of peer effects. Formally, I performed the Pearson test for each school and, under the assumption that schools in Israel are independent, I also added up the test statistics for all schools to obtain an aggregate test statistic such as that described by DeGroot (1984). Obviously, I performed this test only on the basis of the subsample of schools (482 out of 605) that had two or more classes within the relevant grade. Of the 1928 p values for primary schools, 83—only 4% of the sample—were lower than 5%. Furthermore, only two schools had

[10] Ministry of Education, Circular from the Director General, March 2000: http://cms.education.gov.il/EducationCMS/applications/mankal/arc//s7bk3_1_8.htm.
[11] See http://cms.education.gov.il/EducationCMS/applications/mankal/arc//sc3ak3_11_9.htm.
[12] See http://cms.education.gov.il/EducationCMS/Units/Sherut/Takanon/Perek7/Chativa/.

two or more of characteristics with p-values equal to or lower than 5%. The aggregate p-values for each of the four characteristics far surpassed 20 %. The middle-school data yielded similar results: of 230 middle schools with two or more classes, 82 were equal or lower than 5% out of 920 p-values. Therefore, in 9% of the cases I cannot reject that there is non-random assignments. However, only in 13 of the 230 schools (exactly 5% of all schools) two or more p-values are equal or lower than 5%. Overall, I conclude that there is no evidence of systematic formation of classrooms with respect to the four measures of student family background measures.

The second question I investigate in this section is whether classrooms that differ in teaching practices differ in pupils' characteristics as well. To test whether classroom teaching practices are statistically independent of each of the student characteristics, I ran balancing tests as are run in randomized trials, using the following OLS regression model:

(3a) $$x_{ics} = \alpha + \phi_{\tau}(TeachingPractice)_{\tau sc} + v_{ics}$$

and the following school fixed effect model:

(3b) $$x_{ics} = \alpha + \theta_S + \phi_{\tau}(TeachingPractice)_{\tau sc} + v_{ics}$$

where θ_S are the school fixed effects. Tables 2 and 3 present estimates from regressions where the dependent variable is a student characteristic and the explanatory variable is one of the four teaching-practice measures. Table 2 presents the results for the fifth-grade sample and Table 3 does the same for the eighth-grade sample. Each table presents the results from two specifications: an OLS regression without any additional control variables and another that includes school fixed effects. I could not add individual fixed effects because the nature of the pupil-background characteristics (father's years of schooling, mother's years of schooling, number of siblings, an indicator of recent immigration, a gender indicator, an indicator of whether a pupil's parents are native Israeli and four other indicators of ethnic origin) generally rules out changes in characteristics between grades 5 and 8. I also ran these balancing tests for class size in search of evidence of selection in allocating school resources to classes.

Tables 2 and 3 offer little evidence for the proposition that students of different family backgrounds are more likely to be in classes that invoke certain teaching practices, conditional on the school they attend. Only eight of the 72 balancing coefficient estimates presented in Table 2 and Table 3 are significantly different from zero at the 10 percent level of statistical significance when school fixed effects are included in the regressions. Half of these imbalance characteristics are with respect to the teaching practices 'transparency, fairness and feedback' and three of them relate to the gender variable. This pattern stands in sharp contrast to the balancing estimates obtained from the OLS equations (without school fixed effects), which yielded 33 estimates that were significantly different from zero. The pattern of selection between schools with respect to parental schooling is negative, meaning that schools with

large enrollments of pupils whose fathers or mothers have many years of schooling have lower intensities of all four teaching practices; ten of the 16 OLS estimates were significantly different from zero. After I added the school fixed effects to the regressions, seven estimates changed signs to positive and, again, only two of these 16 estimates was significantly different from zero. The between- and within-school selection pattern with respect to number of siblings was also mixed, as two of the eight OLS estimates and three of the eight school-fixed-effect estimates were positive. Further, only 2 of the OLS estimates and one of the school fixed effect estimates were significant. Similarly inconsistent is the positive selection pattern in the distribution of teaching practices between schools based on the balancing-test estimates of the proportion of immigrant pupils. I view these inconsistencies in the selection patterns across the various socio-economic proxies as suggestive evidence that even across schools the variation in teaching practices is not an indication of clear meaningful selection that can confound the effect of teaching practices on pupils' academic outcomes in a certain direction. I should emphasize again, however, that once I added school fixed effects to the regressions, all evidence or signs of potential selection disappeared.

This evidence largely confirms that classes in the sample schools are formed randomly within the schools. There is little evidence that students of different family backgrounds are more likely to be grouped in certain classes depending on the school they attend. However, even if classes are formed randomly, they may receive other school resources differentially. For example, if a class ends up with more children from less advantaged family backgrounds purely by chance, the school may assign this class a smaller class size. To shed light on this question, I ran a set of regressions of the teaching-practice variables described in the previous section on class size and its square. The estimates presented in columns 17-18 in Table 2 and Table 3 show that there are no meaningful correlations between class size and any of the four teaching-practice variables.

5. Results

I now analyze the effects of the various measures of teachers' teaching practices on students' test scores. Table 4 presents descriptive statistics for the four teaching-practice measures. In the Appendix, I present the items that I averaged into each of these indices. Even though each of the items ranges from 1 to 6, the range is narrowed when aggregated into the four teaching-practice measures, from about 2–3 to about 5–6. There are no significant differences between the descriptive statistics of the panel and the full samples.

Table 5 reports estimates based on pooling of all four subjects together and each specification in the table includes subject fixed-effect indicators. I report the effect of each category of teaching practice. Although estimates for all individual items that I used to construct the aggregate teaching-practice measure are not reported here, I should note that most of these estimates are not precisely measured and

some have negative values, partly because of the high correlation among the various items. Therefore, it is appropriate and necessary to aggregate the items in several principal components as I do in this paper. Having no prior information with which to justify a particular weighting, I assign equal weight to all items grouped in a given teaching-style characteristic in order to provide a more transparent interpretation. I computed the class-level mean of these teaching-practice characteristics for each student while excluding the student's own answer. When I used measures based on means that included also students' own answers, I obtained the same results exactly. I also tried averaging the z scores of each item instead of the absolute value of the students' response to each of the questions and I obtained again the same results. I tried as well to avoid the cardinality assumption involved in the use categorical response of students by using instead (0/1) dummy variables that divide to two groups only the class means. The results from theses ordinal measures of teaching practices are fully consistent with the results from the cardinal measures and I will provide more details on this issue later in this section.

The first and second columns in Table 5 report OLS estimates from an equation that included the following control variables: pupil's background characteristics (gender, father's and mother's years of schooling, number of siblings, an indicator of immigration status, and five indicators of ethnic origin—Europe/America, Asia/NorthAfrica, Soviet Union, Ethiopia, and Israel). I also included in the regression as controls the class-level means of all these indicators, an indicator for each of the four subjects, and measures of class climate and learning environment such as level of noise, disturbances, violence, and lack of discipline. In Column 1, the estimates are from separate regressions in which each teaching practice enters as a single treatment variable. In Column 2, the estimates come from one regression that includes all the teaching practice measures as multiple treatments. In Columns 3 and 4, I omit the set of individual characteristics and include individual fixed effects instead. In the specification presented in Columns 5 and 6, I also include primary- and middle-school fixed effects in addition to the controls included in the specification of Columns 3–4. In Panel A, the results are based on a sample of schools that have at least five pupils in the panel data (the Five plus sample); in Panel B, the sample is further restricted to include schools that have at least ten pupils in the panel data (the Ten plus sample). The estimates in Panel B provide a robustness check to the sensitivity of the estimates to the precision of the school-fixed-effect estimates.

Focusing first on the OLS estimates of the effects of teachers' pedagogical methods, we see that all eight estimates are positive but only two (the estimated coefficients of T1) are significantly different from zero. This suggests that there is no obvious selection-bias pattern for three of the treatment measures even if based on the OLS estimates. It is also noticeable that the Column 1 and Column 2 estimates show no clear differences in sign and precision. However, adding of a pupil fixed effect to the equation induces a major change in the size and sign of the estimates of the teaching-practice measures.

Focusing on the specification that includes all treatment measures in the regression, we observe the following: the estimate of T1 (Instilment of Knowledge and Enhancement of Comprehension) drops from 0.381 to practically zero, 0.013. The estimate of T2 (Instilment of Analytical and Critical Skills) climbs from zero to 0.100 (se=0.043) relative to its respective OLS estimate. The estimate of T3 is reversed in sign and becomes significant at –0.084 (standard error=0.031) and the estimate of T4 (Transparency, Fairness, and Feedback) is almost unchanged and still positive and significant. Adding the primary- and middle-school fixed effects to the regression leads to an increase of the point estimates of T1 from zero to 0.144 (se=0.058) while the estimate of T2 is unchanged at 0.099 (se=0.055). The other two estimates (of teaching practices T3 and T4) are very small and not statistically different from zero.

The estimates reported in Columns 5–6 have two remarkable features. The first is the similarity between the estimates when only one of the teaching-practice measures is used as a treatment measure and when all four are used jointly. This pattern suggests that there is very little omitted variable bias when three of the four teaching-practice measures are left out of the equation once we include pupil and school fixed effects. By implication, if it is selection that generates the results about the positive effect of T1 and T2, it must be very different for each of these two measures, which is very unlikely. The second noteworthy feature of the estimates in Column 6 is the different sensitivity of the estimates of T1 and T2 to the addition of school fixed effects to the regressions. While the estimate of T2 remained unchanged at 0.10, the estimate of T1 went up from 0.013 to 0.144. This suggests that while adding a pupil fixed effect is enough to cleanse the T2 estimate of bias due to selection in the distribution of teaching styles across schools and classes, such is not the case for estimating an unbiased estimate of T1, because in this case school fixed effects play a major role in accounting for the selection bias. I should also note that the estimates of the other two teaching practice measures (T3 and T4) also change drastically when school fixed effects are added, from large and significantly different from zero to much smaller and imprecise.

Panel B of Table 5 presents estimates based on a sample restricted to schools that had at least ten pupils in the panel data. The estimated effects do not change at all in comparison to those presented in Panel A. This restriction allows greater precision in estimating school fixed effects and it is important that the estimated effects of T1 and of T3–T4 are not sensitive to this sample restriction.

Before discussing the effect size of the various estimates, it is important to rule out the possibility that if there is a correlation between specific teaching practices and overall teacher effectiveness, our findings then simply reflect the underlying teacher effectiveness and not the effect of the practice itself. For example, if a teacher with strong thinking skills is more effective than one with weak thinking skills, using either “modern” or “traditional” practices, we would expected to see the patterns reported above. However, this does not seem to be driving our results as demonstrated by the very different estimated

effects of T2 and T3. Even though these two teaching practice are highly correlated (estimated correlation coefficient of 0.87, see Table A1) and both are integral elements of modern teaching, T2 has a significant effect on test score while T3 has a zero effect. If it is overall teachers effectiveness and not a specific teacher practice that improve student outcomes, we should have observed T3 being equally effective as T2. I will return to this issue in the next section when presenting evidence about the heterogeneity of the effect of T1 and T2 across sub-groups of students, by gender and by ability. We would not expect such treatment heterogeneity if it is the overall teacher's effectiveness that is causing students' value added instead of some specific teaching practices.

Overall, the evidence in Table 5 strongly suggests that two of the four teaching styles and methods have positive and meaningful effects on pupils' learning. The more important of them in terms of effect size is the indicator of the extent to which teachers make sure that their students know and understand the material by using examples, memorization techniques, homework, classwork, and so on. When the minimum (value 2.5) of this teaching-practice measure rises to the maximum (value 6.0) observed in the data (in the full sample), the test score changes by an average of 0.50 (3.5 * 0.144) standard deviation of the test-score distribution. The effect of elevating teachers' instilment of analytical and critical skills in the classroom from the minimum (2.0) to the maximum (6.0) observed in the data (full sample) is smaller at "only" 0.40 (4 * 0.10) of a standard deviation. If the intensity of the two teaching practices is raised to the maximum level observed in the data relative to a teacher who uses neither of them (measured by the minimum observed in the data), pupils' test scores rise almost one (0.90) standard deviation. A more realistic simulation is to compute the effect size of a change in the intensity of each of the two teaching practices from the mean in the sample (4.2 for T1 and 3.3 for T2) to the best possible (6 for T1 and 5.8 for T2). This simultaneous change improves the average test score in each subject by 0.53 (1.8*0.144+2.7*0.10) of a standard deviation—much more than the effect of most other effective interventions, such as reducing class size, increasing schooling time, or providing teachers and students with conditional financial incentives. It is also probably less expensive to train teachers in the appropriate use of the two teaching practices that I estimate very effective than to apply the other interventions listed. Before discussing the policy implications of the findings, however—I do this in the Conclusions section—I present additional evidence about the heterogeneity of the effect of T1 and T2.

Results Based on Ordinal Measures of Teaching Practices

I have treated so far the teaching practices variables as cardinal treatment. Since these measures are based on categorical responses, there is no obvious justification for treating the numerical values assigned to the categorical responses as meaningful. To assess how sensitive are the results to the cardinal measurement, I also estimated models with the two ordinal measures of the treatment variables

that I described in the data section. I first present In Table 6 estimates based on measures where the treatment variables are dummies of high and low class averages of the regular treatment, where the threshold changes from the 50th to the 90th percentile. Each column presents estimate from one regression using the specification with school in pupil fixed effects. The first column present estimates when the dummy indicators are based on the median, and then the cutoff point increases until the 90th percentile (in column 5). The estimated effect of T1 is positive and significantly different from zero in all five columns while the estimated effect of T2 is also positive throughout but it is large and precisely measure only when the cutoff is the 80th or higher percentile. This is an indication of a non-linear pattern in the effect of T1 and T2, which I will explore below further.

In Table 7, I present estimates of the effect of teaching practices where I measure the intensity of each teaching practice by the proportions of answers above a certain level (4, 5 or 6) in all the questions that the teaching practice consist of. Column 1 presents estimates where the proportion count includes all answers from 4 and up (students who said that part or large part or almost all of the teachers apply the relevant teaching practice). The proportion of students for this level of treatment of T1 is 0.816 and for T2 it is 56.1 The Effect of T1 and T2 are positive and significantly different from zero and the former is larger. The effect of T3 and T4 are practically zero. The Evidence presented in columns 2 and 3 depicts the same pattern which suggest that T1 has a large effect even when only part of the teachers are practicing T1 while the effect of T2 becomes very large when almost all teachers practice it in class.

Overall, the results of Tables 6 and 7 are consistent with the evidence presented in the last column of Table 5. This consistency suggests that cardinality is not a limitation of the measure used in Table 5 and therefore I will continue using it in the rest of the paper.

Estimated Effects by Subject

The results reported so far assume that the effect of each of the teaching practices is the same in all subjects. To test this assumption, columns 1–2 in Table 8 present evidence based on the pooling of math and science test scores and columns 3–4 do the same on the basis of pooled test scores in Hebrew and English. All estimates in columns 1–4 of Table 8 are predicated on regression specifications that include pupil and primary- and middle-school fixed effects. I view these groupings as less restrictive than the pooling of all four subjects together because pedagogy and teaching style may well be more similar in the first two subjects, which are intensive in math and rigorous analysis, and in the two language subjects. Comparison of the estimates in Columns 2 and 4 reveals a surprisingly close similarity in the respective estimates of the effect of all four teaching practices. Based on the estimates in Panel A, the effect of instilment of knowledge/comprehension measure (T1) is 0.141 for math and science and 0.149 for Hebrew and English. The estimate of enhancement of analytical and critical skills (T2) is 0.066 for

math and science and 0.118 for Hebrew and English. The estimates for the other two teaching style measures are small, practically zero, and insignificant for both sets of subjects. Note also the consistency in sign of these two measures in both sets of subjects: T3 is negative in math and science and in Hebrew and English, and T4 is positive in both. The results presented in Panel B strengthen these conclusions. In fact, in the 10+ sample the effect of T2 in math and science is even more similar to that in Hebrew and English (0.089 and 0.118, respectively). This suggests that pooling all four subjects in estimation is not overly restrictive and offers the advantage of estimating the parameter of interest more precisely.

In Table A2, I present evidence based on a separate regression for each subject. The four estimates of T1 are very similar, highest for English (0.173) and lowest for Hebrew (0.120). However, running separate regressions for each subject comes at the expense of precision of estimates because of the smaller sample size used in each regression. The estimates of T2 are not very different for English, Hebrew and science but it is much lower in math.

Non-Linear Effects

The models estimated in Table 5 assumed that the various teaching-practice measures have a linear effect. Table 9 presents results from a specification that allows the effect of each teaching-style measure to be different at low, medium, and high levels of treatment. I divided the distance from the minimum to the maximum value of each teaching measure into three equal segments and allowed the effect of each treatment to vary by these segments as follows:

$$(4) \qquad y_{icst} = \alpha_{ics} + \beta_s + \gamma_t + S_{cs}^{'}\lambda_2 + \sum_{\tau}\sum_{j}\theta_{\tau j}(TeachingPractice)_{\tau sc}q_j + \varepsilon_{icst} + \in_{cst}$$

where j is an index of the treatment intensity, and q_j is a (0/1) dummy indicator for the three possible levels of treatment intensity, low (lowest third), middle (middle third) and highest (highest third).

The first column in Table 9 presents the estimated main effect of each teaching practice; the second and third column estimate the interaction of each teaching measure with dummy indicators of its middle and high segments. The estimates in Row 1 of Table 9 suggest clearly that the gain in test scores due to an increase in T1 is largest when this measure is changed from very low to about its middle range, although the differences are neither statistically significant nor very large. For example, the effect due to changes in the upper range of T1 is just 40% lower than the effect at low values of T1.

The pattern of the non-linear effect of T2 is remarkably different from that of T1 in that its effect is highest in the upper third of its distribution. The estimated coefficient in the upper segment is 0.266 (se=0.117), much lower in the middle range (0.067), and practically zero in the low segment. This suggests that the enhancement of "traditional"-style teaching, based on memorization and repetition with a focus on making sure that all students understand the material taught in class, has a high payoff even at

low intensity treatment, while a teaching style that emphasizes analytical skills yields returns only when used intensively in class. Note that the patterns of the respective estimates for the other two teaching practices suggest no meaningful and significant effect on students' achievements, as was evident in the linear-effect model.

Heterogeneous Treatment Effects

To gain further insights on the extent of effects of teaching styles and methods on students' test scores, I explore heterogeneous effects across different dimensions. Table 10 reports heterogeneous treatment effects of the teaching-style measures by gender and by father's years of schooling (above or equal to/below the median, i.e., 13 years).[13] I prefer to stratify the sample by these subgroups instead of using interaction terms for these subgroups with the treatment effects because in the latter approach the treatment-interaction terms may pick up variations by gender or parental schooling in the effects of other covariates included in the regressions. The stratifying approach comes at the price of estimating the heterogeneous treatment effects on the basis of a smaller sample. Since the sample-size issue is especially troubling in estimating fixed-effects models, Panel A of Table 10 shows the results from samples that are restricted in each subgroup to the inclusion of at least five pupils per school.

In Columns 1 and 2 of Table 10, I report the effect of each of the four teaching-practice measures on boys and girls, respectively. The estimates of T1 ("traditional" teaching) presented in the first row reveal a striking difference in the effect of this teaching practice on boys and on girls. The effect on boys is not statistically different from zero (0.031, se= 0.091) while the estimated effect for girls is very large, 0.237, and precisely measured (se=0.086). The estimated standard errors of these parameters clearly allow us to accept the hypothesis that the practice is more effective for girls than for boys. The effect of T2 ("modern" teaching), on the other hand, is not differentiated by gender. Although the point estimate for boys is larger than that of girls, the precision of both estimates does not allow us to reject the proposition that they are equal. However, another gender difference is seen in the estimates of the transparency, fairness, and feedback practice (T4), which are positive and significant for boys and small, negative, and not different from zero for girls. The implications of this result—that solving problems and exercises routinely in class and teaching based on repetition of material until most students attain knowledge and comprehension affect girls in the main, as opposed to boys—are important and interesting. First, they are consistent with several studies, which show that other schooling interventions either affect only girls (e.g.., pupil monetary incentives) or affect both genders equally (e.g., teacher financial incentives or

[13] Students with missing values in parental education (4% of the total sample) are excluded from this analysis. The results are not sensitive to the inclusion of these students in the low or high education group. Estimating heterogeneous treatment effects by mother's schooling, the results obtained are very similar to those based on father's schooling.

class-size reduction). Second, this heterogeneity in treatment effect is evidence that the effect of T1 and T2 are not simply capturing the overall teacher's effectiveness because in that case we would expect to observe each teaching practice to affect similarly boys and girls.

Next, I stratified the sample by high or low father's years of schooling. Since the father's schooling variable has fewer missing values, I used it to define this indicator even though the evidence based on mother's years of schooling is very similar. Overall, there is not much heterogeneity between the groups in the effect of T1 and T2. The effect of T1 on pupils of low socioeconomic status (SES), while larger, is not statistically different from the effect on high-SES pupils. Yet the economically meaningful higher effect of T1 on low-SES pupils is intuitive since this teaching style, which emphasizes practice and rote learning in the classroom, most likely replaces relatively scarce home and parental guidance in the production of knowledge among low-SES families. However, the similarity in the effect of the modern style of teaching on the two groups may be unexpected but is encouraging because it provides an important indication that pupils from low SES-backgrounds can be equally motivated and challenged by a teaching style that emphasizes the instilment of learning skills and creative thinking. It would be useful to stratify the samples of low and high SES by gender as well in order to gain more insights about the heterogeneity of these two very quantitatively important teaching practices in the classroom, but it is not possible in this study due to sample-size limitations.

Table 11 presents evidence about the heterogeneous effect of T1 and T2 on students' ability as measured by their ranking in the test-score distribution. I defined a new outcome measure as a product of the z score and the 1/0 indicator based on the percentile ranking of the pupil's test score in a given subject. The table offers estimates for five regressions: the 25^{th}, $50^{th,}$ 75^{th}, 80^{th} and 90^{th} percentiles. The pattern of the results based on these percentiles is a good representation of the evidence obtained for other percentiles. Clearly, instilment of knowledge and enhancement of comprehension is very effective in the classroom for pupils below median ability; its effect drops sharply after this threshold is surpassed. In contrast, the effect of instilment of analytical and critical skills is not very effective at very low ability (below the 25^{th} percentile) but its effect picks up and remains high until the very high level of ability, although it is highest at around the 75^{th} percentile. However, the differences in the effect across the distribution of ability levels above the first quartile are not large and sharp enough to draw firm conclusions about such heterogeneity. The results in Table 11 are consistent with the evidence presented in Columns 3–4 of Table 10 due to the positive and large correlation that exists between ability (ranked by test scores) and socioeconomic background.

The evidence about the heterogeneous treatment effects of T1 and T2 is important because it adds credibility to the causal interpretation of the estimates. I showed in Section 4 that classes within schools are formed randomly with respect to parental schooling and student's gender; therefore, any

unaccounted-for sorting or selection of teaching-practice measures across classes within schools should not be different by gender or by parental schooling. The evidence of differential gender and parental-schooling effects is an indication that potential omitted selection or sorting factors, as well as the possibility of endogeneity of the teaching practices, cannot account for the results I present in this paper. The heterogeneity in treatment effects is also an indication that the estimates do not trace to a pattern in which students prefer the teaching practices that their favorite teachers use. This conclusion is enhanced by the way I computed the measures of teaching practices—averages based on the assessment of all students in the class—and by the insensitivity of the results to excluding or including own assessment in computing the class average. Finally, the heterogeneity in treatment effects is also an indication that the estimates do not reflect overall teacher effectiveness. If it is the latter, we would have expected to see similar effects for boys and girls, and also for high and low ability students.

6. Conclusions

In this paper, I measured empirically the relationship between classroom teaching practices and student achievements. I found very strong evidence that two important teaching practices cause student achievement growth. In particular, classroom teaching that emphasizes "instilment of knowledge and comprehension" has a very strong and positive effect on test scores, especially of girls and of pupils from low socioeconomic backgrounds. Second, the use of classroom techniques that endow pupils with "analytical and critical skills" has a very high payoff, especially among pupils from educated families. Transparency in the evaluation of pupils, proper and timely feedback to students, and fairness in assessing pupils also lead to cognitive achievement gains, especially among boys. However, "instilment of the capacity for individual study" does not cause any gain in value added of students learning. These results are robust to the three different methods I use to aggregate the students' responses about their teachers' teaching practices to class level means.

Beyond estimating models with student and school fixed effects that control for all sorts of selection and sorting of students into schools, this paper provides several additional reasons to believe that selection and sorting are not responsible for the results summarized above. I show, for example, that boys and girls were sorted similarly across schools; thus, if the results trace to sorting, they should have been similar, rather than very different, for males and females. In addition, there does not appear to be any sorting within the sample by parental schooling yet the effects of treatment are again heterogeneous. Similarly, if the effect of teaching practices is simply capturing overall teacher effectiveness, we would not expect to see heterogeneity in effect by teaching practices and by gender ability groups of students, but we do. All this evidence supports a causal interpretation of the estimated effects summarized above.

The evidence I presented in this paper provides important insights about what does and does not work in the classroom; the set of results is one of the first that clearly identifies actions of teachers that "pay off" versus others that do not. This study has additional policy relevance because its evidence sheds light on the merits of traditional versus modern approaches to teaching, a contrast that has featured in recent policy debates and educational reforms in several countries. This study may be the first to demonstrate that one approach need not crowd out the other and that the two can coexist. The estimated heterogeneity in treatment effects of the two styles and the essence of teaching that I estimated in this paper implies that it is best to target certain teaching practices to relevant customers and also to mix the two in the classroom. However, a limitation of this study that perhaps can be addressed in the future is that teaching practices are measured as a class average and not for individual teachers.

The effect sizes estimated for some of teaching practices are truly impressive, especially relative to the effect sizes of other potential interventions such as reducing class size, increasing school hours of instruction, and providing more teacher training. These three alternative interventions and other possible educational programs are much more expensive or difficult to implement than the installation of appropriate teaching practices in the classroom. Although this change would entail some teacher training, it should not be too costly since teachers in most education systems around the world routinely engage in on-the-job training. Therefore, re-directing the syllabus relating to enhancement of teachers' human capital toward training in adequate use of "instilment of knowledge and enhancement of comprehension" and "instilment of analytical and critical skills" should be neither too difficult nor too costly. The potential gains seem enormous and worth the effort to sway away teachers from teaching practices that suit their comparative advantages but may not be effective.

7. References

Ammermueller, Andreas and Jorn-Steffen Pischke, (2009). "Peer Effects in European Primary Schools: Evidence from PIRLS," *Journal of Labor Economics*, vol. 27, no. 3, pp. 315–348.

Aaronson, Daniel, Lisa Barrow and William Sander (2007). "Teachers and Student Achievement in Chicago Public High Schools," *Journal of Labor Economics* Vol. 24, No. 1, pp. 95–135.

Bill and Melinda Gates foundation. 2010. "Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project" MET project, Research Paper, December.

Bloom, Benjamin S. (ed.), 1956, Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain, New York, N.Y: Longmans, Green and co.

Cantrell Steven, Kane, J. Thomas, Jon Fullerton, and Douglas Staiger. "National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment," NBER Working Paper #14608, December 2008.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, AND Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project Star", 2010, forthcoming, *Quarterly Journal of Economics.*

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L. "Teacher-Student Matching and the Assessment of Teacher Effectiveness" *Journal of Human Resources* 41(4):778–820 (2006).

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L. (2007). "How and Why Do Teacher Credentials Matter for Student Achievement?" NBER Working Paper 12828.

DeGroot, Morris. 1984. Probability and statistics. 2nd ed. Reading, MA: Addison-Wesley.

Goldhaber, Dan. 2007. "Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness?" *Journal of Human Resources* 42(4): 765–94.

Gordon, Robert, Thomas J. Kane and Douglas O. Staiger, (2006). "Identifying Effective Teachers Using Performance on the Job" Hamilton Project Discussion Paper, Published by the Brookings Institution.

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

Hanushek, Eric A., John F. Kain, Steven G. Rivkin, "Why Public Schools Lose Teachers", *Journal of Human Resources* 39(2), Spring 2004b.

Hanushek, Eric A. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature*, September 1986, 24(3), pp. 1141–1117.

Harris, Douglas, and Tim R. Sass. 2006. "Value-Added Models and the Measurement of Teacher Quality." Florida State University. Unpublished.

Jacob, Brian and Lars Lefgren (2005). "Principals as Agents: Student Performance Measurement in Education" NBER Working Paper No. 11463.

Jacob, B.A., L. Lefgren, and D. Sims, "The Persistence of Teacher-Induced Learning Gains," NBER working paper #14065, June 2008.

Jakubowski Maciej, "Early tracking and achievement growth" Directorate for Education, OECD, December 2009.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger (2008a). "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." *Economics of Education Review*, 27(6): 615–631.

Kane, Thomas J., Jonah E. Rockoff and Douglas Staiger, (2008b). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation" NBER Working Paper #14607, December.

Kane Thomas J., Eric S. Taylor, John H. Tyler and Amy L. Wooten, "Identifying Effective Classroom Practices Using Student Achievement Data", *Journal of Human Resources*, 2011, 587-613.

Koedel Cory and Julian Betts, 2007. "Re-Examining the Role of Teacher Quality In the Educational Production Function," Working Papers 0708, Department of Economics, University of Missouri.

Konstantopoulos, S. "How Long Do Teacher Effects Persist?" IZA Discussion Paper 2893, 2007.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29(1), 67-101. (2004).

Murnane, R. J. & Phillips, B. R. (1981). "What Do Effective Teachers of Inner-City Children Have in Common?" *Social Science Research*, 10, 83–100.

National Council of Teachers of Mathematics, (1991) "Professional Standards for Teaching Mathematics", Published by the National Council of Teachers of Mathematics.

National Research Council, 1996, "National Science Education Standards", National Academy Press Washington, D.C.

Nye Barbara, Spyros Konstantopoulos and Larry V. Hedges "How Large Are Teacher Effects?", Educational Evaluation and Policy Analysis Fall 2004, Vol. 26, No. 3, pp. 237-257.

Resnick, L. (1987). Education and Learning to Think, Washington D.C.: NationalAcademy Press.

Rivkin, S., Hanushek, E. A. and Kain, J. (2005). "Teachers, Schools, and Academic Achievement," *Econometrica*, 73(2): 417–458.

Rockoff, J. E. (2004). "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94(2): 247–252.

Rockoff, Jonah E., Brian Jacob, Thomas J. Kane, and Douglas O. Staiger (forthcoming). "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy*.

Rockoff, Jonah E., and Cecilia Speroni (2010). "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review*, 100(2): 261–66.
Rockoff, E. Jonah and Douglas Staiger. "Searching for Effective Teachers with Imperfect Information," forthcoming, *Journal of Economic Perspectives*.
Rothstein, Jesse (2010). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, 125(1): 175–214.
Salomon, G. Perkins, D. Theroux, P. (2001). Comparing Traditional Teaching and Student Centered, Collaborative Learning [Online, retrieved 14/6/02] URL: http://shaw.ca/priscillatheroux/collaborative.html.
Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. (2010) "Using Student Performance Data to Identify Effective Teachers and Teaching Practices." *American Economic Review* 100(2):256-260.
Zemelman Steven, Harvey Daniels, and Arthur Hyde, Best Practice, Today's Standards for Teaching and Learning in America's Schools, 1993 and 2005, Heinemann, Reed Elsevier Incorporation.

Appendix A: Questionnaire Items

Instilment of knowledge and enhancement of comprehension

1. The teachers give exercises and assignments that help memorize the material.
2. The teachers ask many questions in class that check whether we know the material well.
3. The teachers commend students who know the material well.
4. The teachers provide many examples that help understand the material.
5. The teachers hold discussions in class that help understand the material.
6. During lessons, the teachers ask many questions that check whether we understand the material well.
7. I understand the teachers' scholastic requirements well.

Instilment of analytical and critical skills

1. The teachers give exercises and assignments whose answers have not been studied in class and are not in the textbooks.
2. The teachers require that we use what we have studied to explain various phenomena.
3. The teachers ask that we find new examples by ourselves for the material we have studied.
4. The teachers ask that we try to find several ways to solve a certain problem.
5. The teachers teach us to find a single common explanation for different phenomena.
6. The teachers give assignments where it is required to analyze material and to relate it to other things we have studied.
7. When there are several ways to solve a problem, the teachers require that we check them all and find the best one.
8. The teachers expect us to ask ourselves whether what we have learned is correct.
9. The teachers teach us how to know whether information we have found is important, relevant, and usable.

Instilment of capacity for individual study

1. The teachers teach us how to learn new topics by ourselves.
2. The teachers require students to utilize many and varied sources of information (newspapers, books, databases etc.).
3. The teachers teach us to observe our environment and to follow phenomena that occur in it.

Transparency, fairness and feedback

1. The teachers explain to me exactly what I have to do to improve my studies.
2. The teachers explain according to what they determine the grades / assessments.
3. The teachers often tell me what my situation is regarding schoolwork.

Individual treatment of students

1. The teachers know what the educational difficulties of each student are.
2. When a student has difficulty with a certain topic, the teachers give him more time to study it.
3. The teachers give homework to every student according to his place in the material.
4. The teachers help every student to learn topics interest him.
5. The teachers give me a feeling that if I make an effort I will succeed more at studies.
6. When a student fails, the teachers encourage him to try again.
7. The teachers always assist me when I need help with studies.

Appendix B: Major Categories in the Taxonomy of Educational Objectives (Bloom, 1956) (http://faculty.washington.edu/krumme/guides/bloom.html)

Categories in the Cognitive Domain: (with Outcome-Illustrating Verbs)

1. Knowledge of terminology; specific facts; ways and means of dealing with specifics (conventions, trends and sequences, classifications and categories, criteria, methodology); universals and abstractions in a field (principles and generalizations, theories and structures): Knowledge is (here) defined as the remembering (recalling) of appropriate, previously learned information.

- defines; describes; enumerates; identifies; labels; lists; matches; names; reads; records; reproduces; selects; states; views.

2. Comprehension: grasping (understanding) the meaning of informational materials.

- classifies; cites; converts; describes; discusses; estimates; explains; generalizes; gives examples; makes sense out of; paraphrases; restates (in own words); summarizes; traces; understands.

3. Application: the use of previously learned information in new and concrete situations to solve problems that have single or best answers.

- acts; administers; articulates; assesses; charts; collects; computes; constructs; contributes; controls; determines; develops; discovers; establishes; extends; implements; includes; informs; instructs; operationalizes; participates; predicts; prepares; preserves; produces; projects; provides; relates; reports; shows; solves; teaches; transfers; uses; utilizes.

4. Analysis: the breaking down of informational materials into their component parts, examining (and trying to understand the organizational structure of) such information to develop divergent conclusions by identifying motives or causes, making inferences, and/or finding evidence to support generalizations.

- breaks down; correlates; diagrams; differentiates; discriminates; distinguishes; focuses; illustrates; infers; limits; outlines; points out; prioritizes; recognizes; separates; subdivides.

5. Synthesis: Creatively or divergently applying prior knowledge and skills to produce a new or original whole.

- adapts; anticipates; categorizes; collaborates; combines; communicates; compares; compiles; composes; contrasts; creates; designs; devises; expresses; facilitates; formulates; generates; incorporates; individualizes; initiates; integrates; intervenes; models; modifies; negotiates; plans; progresses; rearranges; reconstructs; reinforces; reorganizes; revises; structures; substitutes; validates.

Evaluation: Judging the value of material based on personal values/opinions, resulting in an end product, with a given purpose, without real right or wrong answers.

- appraises; compares and contrasts; concludes; criticizes; critiques; decides; defends; interprets; judges; justifies; reframes; supports.

Table 1: Descriptive Statistics and Mean Differences Between the Panel and the Full Sample

	Means, sample of all 5th grade students in 2002	Means, panel sample of 5th grade students in 2002	Means, panel sample of 5th grade students in 2002: schools with at least 5 students in final sample	T-test, differences between means in column 1 and column 2	T-test, differences between means in column 1 and column 3
	(1)	(2)	(3)	(4)	(5)
Characteristics					
Father's years of schooling	12.586	12.735	12.783	0.173 (0.136)	0.223 (0.160)
Mother's years of schooling	12.924	13.089	13.076	0.193 (0.124)	0.173 (0.147)
Number of siblings	2.125	2.004	2.021	-0.141 (0.053)	-0.119 (0.064)
Gender (female=1)	0.498	0.503	0.504	0.006 (0.008)	0.006 (0.009)
Immigration status (immigrant=1)	0.136	0.199	0.186	0.073 (0.018)	0.056 (0.022)
Parents born in Israel	0.525	0.445	0.446	-0.092 (0.021)	-0.089 (0.025)
Schools and pupils					
Number of schools	415	359	122		
Number of pupils	26964	3824	3117		

Notes: Each parameter estimate presented in columns 4 and 5 is obtained from a separate regression. Standard deviations are presented in parenthesis in columns 1-3. Standard errors, clustered by school, are presented in parenthesis in columns 4-5 .

Table 2: Balancing Tests, Fifth Grade

	OLS	School FE	OLS	School FE	OLS	School FE
	Father's years of schooling		Mother's years of schooling		Number of siblings	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Instilment of knowledge and enhancement of comprehension	1.350	-0.551	-1.247	-0.438	-0.526	-0.089
	(0.534)	(0.373)	(0.411)	(0.402)	(0.223)	(0.183)
T2: Instilment of analytical and critical skills	-1.750	-0.228	-1.368	0.215	-0.262	-0.265
	(0.470)	(0.381)	(0.401)	(0.345)	(0.186)	(0.141)
T3: Instilment of capacity for individual study	-1.031	-0.151	-0.863	0.074	-0.290	-0.142
	(0.394)	(0.332)	(0.301)	(0.272)	(0.161)	(0.115)
T4: Transparency, fairness and feedback	-1.310	0.322	-1.250	0.080	-0.172	-0.056
	(0.306)	(0.252)	(0.234)	(0.216)	(0.113)	(0.082)
	Recent immigration		Boy		Parents born in Israel	
	(7)	(8)	(9)	(10)	(11)	(12)
T1: Instilment of knowledge and enhancement of comprehension	0.257	-0.031	0.009	-0.053	-0.198	0.018
	(0.064)	(0.049)	(0.037)	(0.068)	(0.078)	(0.070)
T2: Instilment of analytical and critical skills	0.335	0.054	0.005	0.016	-0.272	-0.072
	(0.060)	(0.036)	(0.034)	(0.057)	(0.065)	(0.059)
T3: Instilment of capacity for individual study	0.204	0.030	-0.007	-0.020	-0.182	-0.038
	(0.052)	(0.029)	(0.026)	(0.053)	(0.057)	(0.049)
T4: Transparency, fairness and feedback	0.154	-0.034	0.007	-0.097	-0.172	-0.036
	(0.043)	(0.026)	(0.022)	(0.040)	(0.047)	(0.038)
	Parents born in Asia or Africa		Parents born in Europe or US		Class size	
	(13)	(14)	(15)	(16)	(17)	(18)
T1: Instilment of knowledge and enhancement of comprehension	-0.026	-0.024	-0.035	0.062	-0.707	0.807
	(0.046)	(0.057)	(0.042)	(0.054)	(2.490)	(1.481)
T2: Instilment of analytical and critical skills	-0.007	-0.015	-0.072	0.041	-1.901	0.646
	(0.039)	(0.041)	(0.032)	(0.057)	(1.751)	(0.711)
T3: Instilment of capacity for individual study	-0.013	-0.043	-0.018	0.057	-2.831	0.978
	(0.030)	(0.036)	(0.024)	(0.041)	(1.181)	(0.766)
T4: Transparency, fairness and feedback	0.028	0.026	-0.017	0.056	-1.720	-0.178
	(0.028)	(0.027)	(0.025)	(0.032)	(1.353)	(0.749)

Notes: The table reports OLS and school fixed effects estimates from separate regressions of the relevant dependent variable on each of the four teaching practices. Robust standard errors clustered at the school level are reported in parentheses. The sample includes 5th grade pupils in 2002, from secular schools with 5 or more pupils in the panel data set.

Table 3: Balancing Tests, Eighth Grade

	OLS	School FE	OLS	School FE	OLS	School FE
	Father's years of schooling		Mother's years of schooling		Number of siblings	
	(1)	(2)	(3)	(4)	(5)	(6)
T1: Instilment of knowledge and enhancement of comprehension	-0.399 0.352	0.173 0.254	-0.048 0.357	0.314 0.272	-0.075 0.092	-0.030 0.134
T2: Instilment of analytical and critical skills	-0.637 0.478	-0.299 0.265	-0.381 0.403	-0.310 0.244	0.136 0.114	0.108 0.164
T3: Instilment of capacity for individual study	-0.076 0.322	-0.029 0.145	0.106 0.269	0.004 0.139	0.057 0.073	0.028 0.098
T4: Transparency, fairness and feedback	-0.987 0.299	-0.317 0.181	-0.812 0.282	-0.397 0.209	-0.020 0.077	0.062 0.093
	Recent immigration		Boy		Parents born in Israel	
	(7)	(8)	(9)	(10)	(11)	(12)
T1: Instilment of knowledge and enhancement of comprehension	0.166 0.060	0.040 0.028	-0.023 0.028	-0.076 0.043	-0.143 0.058	0.022 0.037
T2: Instilment of analytical and critical skills	0.174 0.078	0.068 0.049	-0.064 0.038	-0.100 0.051	-0.099 0.078	-0.055 0.042
T3: Instilment of capacity for individual study	0.072 0.054	-0.004 0.028	-0.027 0.027	-0.059 0.037	-0.023 0.059	0.014 0.025
T4: Transparency, fairness and feedback	0.136 0.045	0.029 0.029	-0.068 0.023	-0.129 0.039	-0.166 0.034	-0.021 0.024
	Parents born in Asia or Africa		Parents born in Europe or US		Class size	
	(13)	(14)	(15)	(16)	(17)	(18)
T1: Instilment of knowledge and enhancement of comprehension	-0.045 0.040	-0.049 0.024	0.025 0.022	-0.010 0.027	-0.605 0.958	0.139 1.045
T2: Instilment of analytical and critical skills	-0.070 0.032	0.001 0.038	-0.005 0.034	-0.015 0.033	-2.464 1.282	1.383 1.649
T3: Instilment of capacity for individual study	-0.037 0.028	-0.001 0.021	-0.010 0.020	-0.011 0.021	-2.179 0.986	-0.499 1.026
T4: Transparency, fairness and feedback	0.032 0.032	0.008 0.023	-0.007 0.021	-0.019 0.025	1.296 0.839	0.736 0.873

Notes: The table reports OLS and school fixed effects estimates from separate regressions of the relevant dependent variable on each of the four teaching practices. Robust standard errors clustered at the school level are reported in parentheses. The sample includes 8th grade pupils in 2005, from secular schools with 5 or more pupils in the panel data set.

Table 4: Descriptive Statistics of the Teaching Practices Measures

	Grade 5			Grade 8		
	Mean (1)	Min (2)	Max (3)	Mean (4)	Min (5)	Max (6)
A: Panel sample						
T1:Instilment of knowledge and enhancement of comprehension	4.8 (0.2)	4.0	5.7	4.2 (0.3)	3.2	5.2
T2:Instilment of analytical and critical skills	3.9 (0.3)	3.1	5.3	3.4 (0.2)	2.7	4.2
T3:Instilment of capacity for individual study	3.9 (0.4)	2.9	5.6	3.1 (0.4)	2.1	4.3
T4:Transparency, fairness and feedback	4.2 (0.4)	2.9	5.4	4.1 (0.4)	2.7	5.2
B: Full sample						
T1:Instilment of knowledge and enhancement of comprehension	4.8 (0.3)	3.3	6.0	4.2 (0.3)	2.5	6.0
T2:Instilment of analytical and critical skills	3.9 (0.3)	2.4	5.8	3.3 (0.3)	2.0	5.3
T3:Instilment of capacity for individual study	3.9 (0.4)	1.9	6.0	3.1 (0.4)	1.9	5.5
T4:Transparency, fairness and feedback	4.1 (0.4)	2.3	6.0	4.1 0.4	2.4	6.0

Notes: This table presents means, standard deviations and minimum and maximum values for the four teaching practices measures. Column 1-3 present results for the fifth grade (2002) sample and columns 4-5 for the eight grade (2005) sample. Panel A reports results for the secular schools that are represented in the panel data with 5 or more pupils and panel B reports results for the population of all secular schools. Standard deviations are presented in parentheses.

Table 5: Estimates of the Effect of Teaching Practices on Pupils' Test Scores

	OLS		Pupil fixed effect		Pupil and school fixed effect	
	Each measure included separately	All measures included jointly	Each measure included separately	All measures included jointly	Each measure included separately	All measures included jointly
	(1)	(2)	(3)	(4)	(5)	(6)
A. Five plus sample (N= 20666)						
T1: Instilment of knowledge and enhancement of comprehension	0.288 (0.111)	0.381 (0.129)	0.035 (0.036)	0.013 (0.041)	0.206 (0.049)	0.144 (0.058)
T2: Instilment of analytical and critical skills	0.040 (0.089)	0.003 (0.120)	0.029 (0.032)	0.100 (0.043)	0.146 (0.043)	0.099 (0.055)
T3: Instilment of capacity for individual study	0.006 (0.069)	0.076 (0.093)	-0.058 (0.024)	-0.084 (0.031)	0.043 (0.031)	-0.046 (0.041)
T4: Transparency, fairness and feedback	0.049 (0.062)	0.084 (0.068)	0.078 (0.023)	0.113 (0.026)	0.091 (0.032)	0.028 0.037
B. Ten plus sample (N= 18168)						
T1: Instilment of knowledge and enhancement of comprehension	0.363 (0.124)	0.429 (0.142)	0.012 (0.039)	-0.005 (0.044)	0.221 (0.052)	0.164 (0.061)
T2: Instilment of analytical and critical skills	0.060 (0.104)	-0.015 (0.135)	-0.018 (0.035)	0.039 (0.046)	0.147 (0.046)	0.111 (0.059)
T3: Instilment of capacity for individual study	0.026 (0.076)	0.074 (0.100)	-0.071 (0.026)	-0.079 (0.032)	0.024 (0.033)	-0.076 (0.043)
T4: Transparency, fairness and feedback	0.086 (0.070)	0.094 (0.076)	0.084 (0.026)	0.118 (0.029)	0.092 (0.035)	0.026 (0.040)

Notes: This table reports OLS (columns 1-2), pupil fixed effects (columns 3-4) and pupil and school fixed effects(columns 5-6) estimates of the effect of the four teaching practice measures on pupils test scores measured as z scores. The z scores are computed based on the full sample. For the OLS regressions the standard errors are clustered at the school level and robust standard errors are reported for the pupil and pupil and school fixed effects regressions. The OLS regressions include as controls pupils' personal characteristics (parental education, gender, number of siblings, immigrant status and three ethnic indicators), four subject dummies, class size, class means of the pupils' characteristics and several class climate measure (class means of level of noise, incidence of violence, discipline). The estimates presented in the odd columns are from regressions when each of the teaching practices is used as the only treatment variable in the regression. The estimates presented in the even columns are from regressions where all four teaching practices measures are used simultaneously as treatment variables in the regressions. The estimates presented in panel A are based on the five plus sample and the estimates in panel B are based on the ten plus sample.

Table 6: Estimates of the Effect of Ordinal Measures Teaching Practices on Pupils' Test Scores

	50th percentile	60th percentile	70th percentile	80th percentil	90th percentile
	(1)	(2)	(3)	(4)	(5)
T1: Instilment of knowledge and	0.078 (0.023)	0.048 (0.025)	0.077 (0.026)	0.035 (0.029)	0.075 (0.040)
T2: Instilment of analytical and critical	0.023 0.022	0.036 (0.023)	0.024 0.024	0.080 (0.030)	0.109 (0.042)
T3: Instilment of capacity for individual study	0.011 (0.021)	0.004 (0.022)	0.001 (0.022)	-0.053 (0.032)	0.026 (0.040)
T4: Transparency, fairness and feedback	-0.025 (0.022)	-0.011 (0.022)	0.038 (0.023)	0.039 (0.024)	0.071 (0.031)
N	20660				

Note: The estimates reported in each column of the table are obtained from one regression. The treatment variables are dummies of high and low class averages of the regular treatment, where the cut moves from the 50th to the 90th percentile. The regression specification includes pupil and school fixed effects. See note to Table 5 for additional controls included in the regression. The regressions are based on the five plus sample.

Table 7: Estimates of Effect of Alternative Ordinal Measures of Teaching Practices on Pupils' Test Scores

	Proportion of answers 4 and above	Proportion of answers 5 and above	Proportion of answers 6 and above
	(1)	(2)	(3)
T1: Instilment of knowledge and enhancement of	0.668	0.352	0.138
	(0.195)	(0.173)	(0.220)
	[0.816]	[0.556]	[0.257]
T2: Instilment of analytical and critical skills	0.363	0.289	0.965
	(0.193)	(0.217)	(0.316)
	[0.561]	[0.332]	[0.141]
T3: Instilment of capacity for individual study	-0.106	-0.191	-0.022
	(0.145)	(0.159)	(0.211)
	[0.532]	[0.318]	[0.136]
T4: Transparency, fairness and feedback	0.098	0.198	0.147
	(0.138)	(0.126)	(0.158)
	[0.674]	[0.488]	[0.265]
N	20660		

Note: The estimates reported in each column in the table are obtained from one regression. The treatment variables are proportions of answers above a certain level (4, 5, 0r 6) in all the questions tha the treatment consist of. The figures in square parenthesis are the average proportions in the panel. The regression specification includes pupil and school fixed effects. See note to Table 5 for additional controls included in the regression. The regression is based on the five plus sample.

Table 8: Effects of Teaching Practices by Pooled Math and Science and pooled Hebrew and English Samples, Based on a Model with Pupil and School Fixed Effects

	Math and Science		Hebrew and English	
	Each measure included separately	All measures included jointly	Each measure included separately	All measures included jointly
	(1)	(2)	(3)	(4)
A. Five plus sample				
T1: Instilment of knowledge and enhancement of comprehension	0.180	0.141	0.232	0.149
	(0.066)	(0.078)	(0.068)	(0.080)
T2: Instilment of analytical and critical skills	0.110	0.066	0.171	0.118
	(0.059)	(0.075)	(0.059)	(0.077)
T3: Instilment of capacity for individual study	0.032	-0.037	0.051	-0.058
	(0.042)	(0.055)	(0.044)	(0.057)
T4: Transparency, fairness and feedback	0.067	0.013	0.116	0.044
	(0.044)	(0.051)	(0.045)	(0.052)
N	10464	10464	10196	10196
B. Ten plus sample				
T1: Instilment of knowledge and enhancement of comprehension	0.193	0.177	0.250	0.155
	(0.071)	(0.083)	(0.073)	(0.085)
T2: Instilment of analytical and critical skills	0.107	0.089	0.175	0.117
	(0.063)	(0.080)	(0.063)	(0.081)
T3: Instilment of capacity for individual study	0.003	-0.073	0.042	-0.082
	(0.045)	(0.058)	(0.047)	(0.060)
T4: Transparency, fairness and feedback	0.048	-0.008	0.138	0.061
	(0.048)	(0.054)	(0.048)	(0.055)
N	9178	9178	8990	8990

Notes: Columns 1-2 report the estimates and robust standard errors (in parentheses) from regressions of the four treatment variables on pupils' achievements based on a sample that includes the Math and Science test scores. Columns 3-4 report the respective estimates from a sample that includes the Hebrew and English test scores. The estimates are based on the pupil and school fixed effect model. See notes to Table 5 for the controls included in these regressions. The estimates presented in panel A are based on the five plus sample and the estimates in panel B are based on the ten plus sample.

Table 9: Estimates of Non Linear Effects of Teaching Practices

	Main Effect	Main Effect Interacted with a dummy indicator of middle third for the teaching practice distribution	Main Effect Interacted with a dummy indicator of the highest third of the teaching practice distribution
	(1)	(2)	(3)
T1: Instilment of knowledge and enhancement of comprehension	0.199 (0.069)	-0.041 (0.039)	-0.073 (0.056)
T2: Instilment of analytical and critical skills	0.011 (0.064)	0.067 (0.034)	0.266 (0.117)
T3: Instilment of capacity for individual study	-0.052 (0.044)	0.026 (0.034)	0.395 (0.358)
T4: Transparency, fairness and feedback	0.003 (0.049)	0.041 (0.043)	0.053 (0.060)
N	20660		

Notes: The estimates reported in this table are obtained from one regression. The distribution of each teaching practice measure was divided to quintiles. The first column reports the main effect of each teaching practice measure. The second and third columns report the estimated effect of the interaction term between each of the teaching practices and each of the dummy indicators of the two upper quintiles. The regression specification includes pupil and school fixed effects. See notes to Table 5 for additional controls included in the regression. Robust standard errors are presented in parenthesis.The regression are based on the five plus sample.

Table 10: Effects of Teaching Practices on Pupils' Test Scores By Gender and Parental Education Based on a Pupil and School Fixed Effect Model

	Boys	Girl	High SES	Low SES
	(1)	(2)	(3)	(4)
A. Five plus sample				
T1: Instilment of knowledge and enhancement of comprehension	0.031	0.237	0.116	0.185
	(0.091)	(0.086)	(0.100)	(0.079)
T2: Instilment of analytical and critical skills	0.134	0.075	0.118	0.103
	(0.090)	(0.080)	(0.101)	(0.075)
T3: Instilment of capacity for individual study	-0.109	-0.059	-0.044	-0.097
	(0.065)	(0.058)	(0.072)	(0.055)
T4: Transparency, fairness and feedback	0.093	-0.041	0.065	-0.003
	(0.058)	(0.056)	(0.064)	(0.052)
N	9398	9428	6546	12038
B. Ten plus sample				
T1: Instilment of knowledge and enhancement of comprehension	0.024	0.189	0.074	0.171
	(0.087)	(0.084)	(0.093)	(0.077)
T2: Instilment of analytical and critical skills	0.117	0.085	0.100	0.094
	(0.086)	(0.078)	(0.095)	(0.073)
T3: Instilment of capacity for individual study	-0.090	-0.037	-0.009	-0.073
	(0.063)	(0.056)	(0.069)	(0.053)
T4: Transparency, fairness and feedback	0.086	-0.036	0.074	0.011
	(0.056)	(0.054)	(0.060)	(0.051)
N	10226	10434	7688	12972

Notes: This table reports pupil and school fixed effects estimates of the effect of the four teaching practice measures on various subsamples of students. All four teaching practices are entered simultaneously as treatment variables in the regressions. The High and low SES samples refer to high and low parental schooling, respectively. Robust standard errors are presented in parenthesis. See notes to Table 5 for list of controls that are included in each of the four regressions.

Table 11: Effect of Teaching Practices by Percentiles of Test Scores Distribution

The outcome measures are the product of the z score and a 0/1 dummy indicator of percentile ranking of students	T1: Instilment of knowledge and enhancement of comprehension (1)	T2: Instilment of analytical and critical skills (2)
Above 25th percentile	0.079	0.022
	(0.031)	(0.029)
Above 50th percentile	0.077	0.046
	(0.036)	(0.034)
Above 75th percentile	0.024	0.065
	(0.033)	(0.031)
Above 80th percentile	0.004	0.053
	(0.030)	(0.029)
Above 90th percentile	0.011	0.043
	(0.024)	(0.023)
N	20660	

Notes: This table reports pupil and school fixed effects estimates of the effect of the first two teaching practice measures on test scores. The two measures are entered jointly as treatment variables in the regressions. The estimates in each column are from one regression based on the five plus sample. The dependant variable is a product of the z_score and the dummy indicator for the specified percentile. The Panel includes pupils from secular schools, with 5 or more pupils in the panel. Robust standard errors are reported in parenthesis.

Table A1: Correlations Between Treatment Variables

	T1	T2	T3	T4
	(1)	(2)	(3)	(4)
T1	1			
T2	0.807	1		
T3	0.808	0.871	1	
T4	0.536	0.446	0.443	1

Notes:
T1: Instilment of knowledge and enhancement of comprehension
T2: Instilment of analytical and critical skills
T3: Instilment of capacity for individual study
T4: Transparency, fairness and feedback

Table A2: Effect of Teaching Practices on Pupils Test Scores by Subjects

	English	Hebrew	Math	Science
	(1)	(2)	(3)	(4)
A. Five plus sample				
T1: Instilment of knowledge and enhancement of comprehension	0.173	0.120	0.128	0.166
	(0.098)	(0.109)	(0.095)	(0.114)
T2: Instilment of analytical and critical skills	0.105	0.150	0.025	0.115
	(0.093)	(0.103)	(0.091)	(0.111)
N	5100	5096	5292	5172
B. Ten plus sample				
T1: Instilment of knowledge and enhancement of comprehension	0.160	0.134	0.168	0.193
	(0.104)	(0.115)	(0.101)	(0.121)
T2: Instilment of analytical and critical skills	0.089	0.167	0.044	0.154
	(0.100)	(0.110)	(0.096)	(0.118)
N	4494	4496	4640	4538

Notes: Columns 1-4 report the estimates and robust standard errors (in parentheses) from regressions of the four treatment variables on pupils' achievements for each subject separately. All four teaching practices are included jointly as treatment variables in the regressions. See notes to Table 5 for the controls included in these regressions. The estimates presented in panel A are based on the five plus sample and the estimates in panel B are based on the ten plus sample.