

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES  
Série des Documents de Travail du CREST  
(Centre de Recherche en Economie et Statistique)

**n° 2008-28**

**Teachers' Training, Class Size  
and Students' Outcomes :  
Learning from Administrative  
Forecasting Mistakes**

**P. BRESSOUX<sup>1</sup> – F. KRAMARZ<sup>2</sup>  
C. PROST<sup>3</sup>**

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

---

<sup>1</sup> Université Pierre Mendès France de Grenoble. Mail : [pascal.bressoux@upmf-grenoble.fr](mailto:pascal.bressoux@upmf-grenoble.fr)

<sup>2</sup> CREST-INSEE. Mail : [kramarz@ensae.fr](mailto:kramarz@ensae.fr)

<sup>3</sup> CREST-INSEE. Mail : [corinne.prost@ensae.fr](mailto:corinne.prost@ensae.fr)

# **TEACHERS' TRAINING, CLASS SIZE AND STUDENTS' OUTCOMES: LEARNING FROM ADMINISTRATIVE FORECASTING MISTAKES\***

Pascal Bressoux, Francis Kramarz and Corinne Prost

This paper uses a feature of the French system in which some novice teachers start their jobs before receiving any training. Moreover, thanks to administrative mistakes in forecasting the number of teachers, trained and untrained novice teachers are similar in 1991. We show that they are assigned to similar classes. In addition, we show that the same sample can be used to estimate the causal effect of class size. Our findings are: (1) teachers' training substantially improves students' test scores in mathematics, except for initially low-achieving students; (2) a small class is beneficial to students, especially to low-achieving ones.

Several papers find a large variability in the extent to which teachers promote student learning.<sup>1</sup> Yet little evidence has been found that any observable teacher characteristic explain a large part of this variation. Teacher experience, at least the very first years of experience, seems to have a small positive impact on student achievement (Rivkin, Hanushek and Kain, 2005; Clotfelter, Ladd and Vigdor, 2006). As for the impact of teachers' academic training, the results are mixed. Some studies show that measures of academic skills are positively related to teacher effectiveness: verbal ability, subject matter preparation, quality of a teacher's undergraduate college, test scores. Yet it seems that teachers holding master's degrees are not more effective (Ehrenberg and Brewer, 1994, 1995; Clotfelter, Ladd and Vigdor, 2006).

Few studies have specified the impact of professional training in developed countries. In the US, some recent papers try to assess the effects of certified programs and those of new programs with reduced requirements prior to teaching (Kane, Rockoff and Staiger, forthcoming; Boyd et al., 2006). Few effects are found and they are at most small.

On the contrary, two papers find a large impact of training in other countries. In France, Bressoux (1996) studies the effect of teachers' training on third-grade pupils' achievement, comparing certified and uncertified teachers. Bressoux finds that training substantially improves students' scores in mathematics. Angrist and Lavy (2001), evaluating the effect of in-service teacher training in Jerusalem schools, find a significantly positive causal effect of this program on pupils' test scores. Their cost-effectiveness analysis suggests that teacher training may provide a less costly means of improving pupil achievement scores than reducing class size or adding school hours.

Importantly though, the experiment used in Bressoux (1996) is not randomized. The ideal situation would involve the random assignment of pupils to the different types of teachers. In fact, Bressoux (1996) uses a specific survey on third-grade students and teachers in 1991, including three types of teachers: untrained novice teachers, trained

---

\* The authors have benefited from helpful comments by participants in the labour economics workshop at Cornell University, the seminar of the Department of Evaluation of the Ministry of Education, the CEPR-EEEPE meeting in Amsterdam, the 2006 ASSA meetings (Boston), the TARGET conference in Vancouver, the INED seminar (Paris), and the 2008 Royal Economic Society Meetings in Warwick. We are particularly grateful to a referee as well as to Ronald Ehrenberg, Caroline Hoxby, Robert Hutchens, George Jakubson, Eric Maurin, and Christopher Taber for their comments and suggestions.

novice teachers, experienced teachers. Yet the data show that classes differ according to the teachers' status. Hence, in the absence of random assignment, Bressoux estimates the impact of training using regressions controlling for numerous variables. The estimated effect is the causal one if no unobserved student or class characteristic is correlated with the teacher's type and with the student's test scores. Otherwise, estimates are potentially biased.

This paper explores more thoroughly the quasi-experimental design of the same data. The strategy relies on the specificity of experienced teachers. The fact that the allocation of classes is not random is virtually only due to experienced teachers, who can choose their schools, and who often choose advantaged zones. It is well known that the effect of teacher experience on student achievement is difficult to assess because teachers are sorted across schools, more experienced teachers often working in schools with affluent students (for a recent paper on this topic, see Clotfelter, Ladd, Vigdor, 2006). But, in principle and in the data, **trained and untrained novice** teachers are assigned to almost similar classes. So our paper uses the fact that, when excluding experienced teachers, we are faced with a quasi-randomized design. Furthermore, because the quality of these trained and untrained novice teachers might differ, our identification strategy relies on administrative mistakes in forecasting the number of teachers, mistakes that appear to have only taken place in the year we study. In addition, and as usual in this literature, we check the robustness of our assumptions on our observable variables.

The data used here are very rich. The unit of observation is the student, a very important element for this kind of analysis (see Summers and Wolfe, 1977). Multiple students' characteristics are collected and students' achievement is extremely precisely measured by detailed test scores at the beginning and at the end of the year. Furthermore, all students within a third-grade classroom are included in the sample. This gives us an opportunity to control for class effects. In addition, teachers provide a lot of information on their personal characteristics, their teaching practices, as well as characteristics of their classes and their schools.

A first aim of this paper is to check that Bressoux's findings on training – better trained teachers induce higher students' outcomes – are robust. To perform this task, we control for the endogenous allocation of classes by excluding experienced teachers, in order to estimate the causal effect of training of novice teachers on pupils. Particular attention is given to heterogeneous effects. A second goal is to see if some particular characteristics of the teachers, such as their university background (which was not used in Bressoux, 1996), have any impact on their students' outcomes. This paper also examines other class characteristics, more particularly class size. Indeed, when excluding experienced teachers, it appears that class size is not correlated with pupils' initial test scores. There is no sign of a relation between class sizes and class mean initial achievement or class socio-economic background. Thus, it seems that no selection bias in class size allocation is present when our sample is restricted to novice teachers. Consequently, we use similar methods to assess the effect of class size as were used to estimate the training effect.

The findings on the training effect are very close to those found by Bressoux (1996): the training of novice teachers largely promotes students' learning in mathematics. Yet it seems that within classes, less able students do not benefit from their teachers' training.

We find that class size has substantial effects on third-grade test scores. These results are close to those of Piketty (2004) on French second-graders, and to those analysing the Tennessee STAR experiment.<sup>2</sup> It appears that the effect is similar on students' reading scores within the classes; it is larger for less able pupils in mathematics. Moreover, a smaller class size has a more favourable impact on students' scores when students are in low-achieving schools, a potential consequence of higher frequencies of disruptions in such schools, as described in Lazear (2001). Furthermore, no apparent relation can be found between class size and teaching practices.

The paper is organized as follows. Following a description of the design and the data in Section 1, Section 2 describes the statistical model and the empirical tests for evidence of non-random differences in the observable characteristics. Section 3 and 4 reports the estimation results for the training effect and the class size effect. Section 5 concludes.

## **1. Institutional Background and Data**

### *1.1. A Quasi-Experimental Design*

In France, teachers in public schools are civil-servants recruited and paid by the State. Recruitment is based on competitive examinations which usually have low acceptance rates (at the end of the 1990's, there were often more than five applicants for each primary school teacher new position). To take the examination, students are required to have a diploma in any academic field, corresponding to at least two years in a university.<sup>3</sup> When the students pass the examination, they become civil servants in traineeship and are trained in specific colleges. After their traineeship, young teachers get tenure. During the period this paper studies, the beginning of the 1990's, the teacher training colleges were called 'écoles normales' and students were trained during two years.<sup>4</sup>

Thus teachers are usually seen as a homogeneous professional group of persons, with equivalent credentials and the same training. Yet, each year, in France at least, some novice teachers start their job without any training.

The number of slots in the training colleges is limited and determined each year by the regional administration, using forecasts for teachers' positions. All applicants are ranked according to their grades in this examination. The students ranked first enter the 'école normale' and begin their training. Students who are ranked just after the last admitted candidate on this primary list are assigned and ranked within a waiting list.

In September, the number of vacant job slots is often greater than the one expected two years earlier. Students who have finished their training at the 'école normale' are assigned to some of these job slots, and, in October, some students in the waiting list are assigned to the vacant slots. Hence, these persons have to teach a class for an entire school-year without receiving any training. They enter the 'école normale' the year after.

This special feature of the French system gives the opportunity to compare teachers who were trained with teachers who were not. Yet it is not obvious that comparing the effectiveness between untrained and trained novice teachers could give an estimate of the impact of training on student achievement. In fact, there are two potential sources of selection bias between trained and untrained novice teachers. The first potential source of bias stems from the fact that trained novice teachers are not exactly similar to untrained teachers: in principle, they have better rankings at the entrance examination than

untrained beginners, who come from the waiting list. This should mean that initial academic knowledge for trained teachers is wider. Yet the rankings of the untrained teachers depend on the year of the exam, and the number of persons directly selected through the primary list. And for any given year (prior to 1992), untrained beginners have taken the exam the same year whereas trained beginners have taken the exam two years earlier. So they were ranked by two different exams taking place in two different years.

The design of this study relies on a specific year, 1991, which ends a series of forecasting errors leading to a large increase in the number of untrained teachers, as can be seen in Figure 1.<sup>5</sup> Moreover, and interestingly for the purpose of this study, in 1991, the number of students selected in the primary list was very small. So the untrained novice teachers, who had taken the entrance examination in 1991, had very good rankings and would have been selected for entry had they competed for the examination during another year, and especially during the year 1989, when the trained novice teachers we examine passed their entrance competitive exam. Thus this type of selection bias is likely to be very small.

Fig. 1

*Number of Individuals Selected or Recruited on the Waiting List*

The second source of bias may come from the usual problem with non- experimental design in education: the assignment of the various types of teachers to classes is non-random and may be related to student achievement. In France, schools do not hire their teachers. Instead, the system of job assignment is centralized and depends on the teachers' choices. When the choices of different teachers are the same, the final assignments depend on the years of experience and on a mark given by the administration, this mark being well correlated with the years of teaching experience. Hence, as they accumulate experience, teachers are able to choose the schools they want, and mostly go from disadvantaged schools to advantaged ones.

On the contrary, novice teachers have much less opportunities to choose their schools and mostly go to schools that have not been chosen by experienced teachers. Nevertheless, there are potentially two reasons of sorting for trained and untrained novice teachers and those reasons would have opposite results. On the one hand, trained novice teachers choose first, so they may choose the best schools. On the other hand, some new slots open between September and October, when jobs are allocated to untrained teachers. These new slots are due to unexpected quits of experienced teachers and can be in good schools (after accidents, diseases, deaths, changes in careers,...). Nevertheless, it should be noted that experienced teachers are not only able to work in the best schools, but are also often allocated to the best classes inside their schools. Finally, it is likely that novice teachers have on average similar classes, being trained or untrained. In the following of the paper, we will test this hypothesis using observables.

## 1.2. *The Data*

The data come from a survey conducted by the French Ministry of Education in the school year 1991-1992, in 12 'départements'. The sample covers third-grade students (8

years old) and their teachers: in French primary schools (1<sup>st</sup> grade to 5<sup>th</sup> grade), students only have one teacher, who teaches all disciplines. The data explicitly includes the three categories of teachers: untrained novice teachers, trained novice teachers and experienced teachers. The sampled teachers were teaching in third-grade classes or in multi-grade classes including third graders. In the 12 'départements' selected, all novice teachers were surveyed while experienced teachers were chosen randomly. All third-graders in the same class were included. Finally, the survey covers 4,001 students and 197 teachers. The numbers of teachers within each category are not perfectly balanced: there are 96 experienced teachers, 65 trained novice teachers and 36 untrained novice teachers.<sup>6</sup>

The information about the students is comprehensive: parents' occupations<sup>7</sup>, sex, date of birth, nationality (French or not), number of siblings, number of years spent in a pre-elementary school, repeated classes. In addition, two sets of scores are available in the data. In France, there is national testing of all pupils just at the beginning of the third grade, both in reading and mathematics. The reading tests comprise grammar, vocabulary, spelling and reading comprehension *per se*. The mathematics tests comprise arithmetic, geometry and problem-solving. For this specific survey, covered pupils were also tested at the end of the school year in both subjects, using a design similar to that prevailing in the entry tests. These both sets of scores are normalized in this paper, with a standard error of 1.

In addition, teachers had to answer a questionnaire on their personal characteristics, on their teaching practices, and on the characteristics of their classes and their schools. The main variables used in the following are the field of specialization of the teacher during his/her studies at the university (sciences, humanities, unknown, other), the class size, the fact that the class is a combination class mixing students from different grades, the category of the area of the school (rural, semi-rural, urban), and the priority status. The class size is not known for 35 classes. Yet when the class size is known and the class comprises only third-graders, the class size variable is equal or close to the number of students at the end of the year, measured by the number of third-graders enrolled in the final test. Thus the unknown class sizes are imputed at this number, except for 8 classes which mix students of different grades. Finally, the mean of the class sizes is 23.9 students per class, with a standard deviation of 4.1. By comparison with data on all elementary schools, Piketty (2004) finds that the average class size in the primary schools (first grade to fifth grade) is close to 23.3 in the school year 1991-1992.

In general, the fraction of missing values is small, except for some variables like test scores where it can amount to 15%. This attrition for test scores is due to many reasons. A few students were present at the beginning of the year, and then moved to another class or school and are not observed at the end of the year. The opposite is true: a few students were not enrolled in the class at the beginning of the year but are present at the end. Moreover, some students are observed in the data but some of their test scores are missing, probably because they were not present in class when the tests were conducted. In addition, in some classes, all the scores are missing. Finally, the scores of initial or final reading tests are not known for 981 students, including 9 entire classes, and the scores of math tests are not known for 893 students, including 11 entire classes.

Each test comprises four exercises. The test score is obtained by adding up the scores of each exercise. Within each test (initial reading, final reading, initial math, final math), the scores of these exercises are very well correlated. We use this result to impute the

total score of those students for whom only one exercise score is missing, through a regression of the total score on the scores of the three other exercises. It is possible to do this imputation for 680 test scores. After this imputation, test scores are missing in 13 entire classes.<sup>8</sup>

Eventually, this attrition and these imputations should not induce any bias: descriptive statistics show that the characteristics of the balanced sample, encompassing students whose scores and class sizes are known or imputed, do not significantly differ from those of all the students. It seems that the absence of information on the scores or the class sizes have random origins.

The data show that the aggregate characteristics of pupils vary with teachers' types (see Table 1). Indeed, experienced teachers have on average better classes. In these classes, compared to those with novice teachers, initial scores are higher, the share of non-French pupils is lower, children have fewer siblings, fathers and mothers have more often a high occupation and students less often repeated the first grade. In addition, schools are less often in a priority educational area.<sup>9</sup> Besides, experienced teachers work more often in large classes. The characteristics of teachers vary with teachers' types as well. Experienced teachers are more often male than novice teachers, they are, of course, older, and they come from families where the parents' socioeconomic backgrounds are lower.

Table 1  
*Statistics by Type of Teacher*

*Notes.* Balanced sample. Standard deviations are reported in parentheses for non-dummy variables. The two last columns give the P-values for the tests of the hypothesis that students are similar across teacher categories. The second to last column compares the three categories whereas the last column compares only students with novice teachers.

By contrast, the classes with trained novice teachers and the classes with untrained novice teachers are essentially similar. The Chi-squared tests do not reject the hypothesis that students are identical whether their teachers are trained or untrained, except for the size of the family. As for school-level variables, trained novice teachers are more often located in urban areas and in priority zones than untrained novice teachers. Finally, teachers' characteristics are also similar. Surprisingly, trained novice teachers are on average the same age as untrained teachers even if we would have expected trained novice teachers to be 2 years older. This is due to a few more untrained teachers who are older than 30.

## 2. Statistical Methods

To assess the trained teachers' effectiveness in raising student achievement, we estimate the following regression at the student level:

$$S_i^f = \alpha S_i^i + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{X}_i^c \boldsymbol{\beta}^c + \delta T_i + \mathbf{Z}_i \boldsymbol{\gamma} + \varepsilon_i, \quad (1)$$

where  $S_i^f$  and  $S_i^i$  are the final and initial test score of student  $i$ ,  $\mathbf{X}_i$  represents student characteristics,  $\mathbf{X}_i^c$  is the characteristics of student  $i$ 's classroom,  $T_i$  is the category of student  $i$ 's teacher and  $\mathbf{Z}_i \boldsymbol{\gamma}$  is the estimation of random class effects. This latter

component allows us to take into account the correlation between students within classes, since class variables may be not sufficient to control for these correlations.

It is important to incorporate class effects; otherwise, the standard deviations could be underestimated, as it would be the case with OLS estimation (Moulton, 1986). However, Moulton stresses the problem of the precision of coefficient estimates, but he also shows that the coefficients may be different when the estimation incorporates random class effects **without** imposing the absence of correlation between these effects and the other covariates. Indeed, this kind of estimation results in substantial gains in efficiency. Throughout this paper, class effects are estimated through the resulting mixed models (Robinson, 1991). These models allow a general specification of class effects, fixed effects being only a specific case within such a specification. In this mixed models setting, identification of class effects uses more information than “classic” fixed effects models -- the variance of the class effects instead of only the mean, thanks to a more general prior distribution.

Bressoux (1996) estimates the training effect through an equation similar to (1). Since experienced teachers have on average better classes, Bressoux (1996) assumes that this selection bias can be controlled for by using observed variables, including initial test scores. The causal interpretation of the coefficients related to the type of teacher relies on the assumption that no selection bias comes from unobserved variables.

This paper explores the issue of this non-randomized design, and assesses the robustness of the teachers’ training effects found in Bressoux (1996). Our strategy is to check the extent of the potential bias using the observable variables. This is done through the regression of initial test scores on student and teacher characteristics:

$$S_i^j = \mathbf{X}_i \tilde{\boldsymbol{\beta}} + \mathbf{X}_i^c \tilde{\boldsymbol{\beta}}^c + \tilde{\delta} T_i + \mathbf{Z} \tilde{\boldsymbol{\gamma}} + \tilde{\varepsilon}_i. \quad (2)$$

The estimated coefficient  $\tilde{\delta}$  shows whether the assignment of classes across the categories of teachers can be considered as random or not. If the coefficient  $\tilde{\delta}$  is significantly different from zero, it means that the assignment is non-random since the students have not been exposed to these teachers’ teaching yet. If the coefficient  $\tilde{\delta}$  is not significant, this is a good sign that the assignment can be considered as random. Yet  $\tilde{\delta}$  can be insignificantly different from zero because of a large standard deviation rather than a small coefficient. Thus the strategy will be to systematically compare  $\tilde{\delta}$  with  $\delta$ , estimated on final scores.<sup>10</sup>

The results of the regression of initial test scores on teachers’ type are given in Table 2. They confirm that experienced teachers teach in better classes. Column 1 reports that the correlation between student initial scores and the indicator variable for teachers’ experience is large and significant, both in reading and in mathematics. These two correlations remain significant, even when controlling for student characteristics (see column 3). On the contrary, it seems that classes with untrained novice teachers and classes with trained novice teachers are not different in terms of initial achievement, since the correlations between initial scores and teacher’s training are not significant, with or without other controls.

Table 2  
*Regression Estimates on Initial Test Scores, with all Pupils*



*Notes.* Regression of the **initial** test scores of **all the students**; when the covariates are included, they comprise student and teacher characteristics. Coefficients are estimated through mixed models, with class effects. Standard errors are reported in parentheses. \*\*p<0.05 \*p<0.10

The robustness of this last result is checked using the same framework, but restricting to the sub-sample of pupils having novice teachers. The results are given in Table 3. The coefficient of the training indicator is never significant.

Table 3  
*Regression Estimates on Initial Test Scores, on the Pupils with Novice Teachers*

*Notes.* Regression of the **initial** test scores of **the students with novice teachers**; when the covariates are included, they comprise student and teacher characteristics. Coefficients are estimated through mixed models, with class effects. Standard errors are reported in parentheses. \*\*p<0.05 \*p<0.10

Our identifying assumption is that trained and untrained novice teachers are randomly assigned to classes, a fact consistent with what was shown using our observed variables. Hence we have chosen to estimate the training effect on the sub-sample of novice teachers. To be able to match classes together, some taught by trained novice teachers when others are taught by untrained novice teachers, we discard those with experienced teachers and focus on pupils with novice teachers. Put differently, we have a sample of similar students, some have trained novice teachers and constitute the treatment group, and some have untrained novice teachers and constitute the control group. Thus, given our previous assessment of the allocation process of novice teachers to classes, including the forecasting mistakes, we can expect that few biases will perturb our estimation and that the resulting coefficient of the training treatment will yield a causal effect. The idea is close to the one used in Angrist and Lavy (2001). In their paper, they observe that pupils in the treatment group have initially lower scores than pupils in the control group. As they would like pupils in the control group to be comparable to pupils in the treatment group, they match individual pupils on the basis of their initial test scores, by dividing test scores into quartiles and comparing treatment and control scores in each quartile. Here, we restrict the sample in order to have similar pupils in the treatment and control group. But, by contrast with Angrist and Lavy, we keep a regression strategy, in order to control for the other covariates, and more specifically to control for class effects.

We will also check the robustness of our results inspiring ourselves from the treatment literature (the treatment is to have a trained teacher) and matching techniques, by focusing on the classes sharing a “common support”. For this, we will estimate a propensity score for a class of having a novice trained teacher, in contrast with having an untrained one. This propensity score will allow us to focus on the common support of our classes.

In the last part of the paper, we apply a similar strategy to estimate the class size effect. Indeed, Table 2 reports that the correlation between initial scores in reading and class size is positive and significant when all pupils are included in the regression. When adding other covariates, this correlation remains significant and positive, even though the effect is weaker. By contrast, Table 3 reports that even without any other control, class size is no more correlated with initial scores when we focus on the reduced sample of students with novice teachers.

Class size is often positively correlated with students' achievement in countries with a centralized school system (see Woessman, 2005; Woessman and West, 2006). In these countries, advantaged schools or districts cannot hire more teachers and the education system is often organized in order to support less advantaged pupils by gathering them in small classes whereas more advantaged students are assigned to larger classes. This can be done within schools as well as between schools, for instance with the implementation of programs which allocate a greater number of teachers to disadvantaged schools. This selection bias is one reason why causal effects of class size can be difficult to measure.

In France, the system of assignment of teachers to schools is centralized, and is not supposed to differentiate between schools in the allocation of resources. The only official exception is the policy of education priority areas (ZEP, 'zones d'éducation prioritaires'). The ZEP policy is a program implemented in 1982, which is supposed to give more resources to disadvantaged schools (for a description and an assessment of this program on sixth and seventh graders, see Benabou, Kramarz, Prost, forthcoming). According to our data, the classes in the ZEP have on average 23.8 students per class, whereas the mean class size in the non priority zones is 25.2.<sup>11</sup>

Rural areas also have schools with smaller classes: because of small enrolments, classes are often small, even if schools are sometimes organized with combination classes by mixing students of different grades in the same classroom. Yet the impact in terms of selection is not clear since pupils in rural schools have better achievement at the beginning of third grade, but tend to improve less during the year.<sup>12</sup>

In addition to these sources of bias, due to sorting between schools in response to the organization of the educational system, biases due to parents' behaviour may also arise. Some parents may move or enrol their children in a private school when their children happen to be enrolled in a large class. If these parents are also those who are the most motivated for their children's education, it would induce a negative bias. Yet we believe that if this bias is large, it should appear through the correlation between class size and initial achievement.

There may also be selection within schools, usually a positive selection; low achieving students being allocated to the smaller classes.<sup>13</sup> This selection is possible in large schools, when there are several third grade classes. Worthy of note, French schools are often small with one class per grade (when the enrolment of third-graders exceeds 30 students, it does not always entail a new third grade class, but some third graders are often assigned to a class with students of other grades).<sup>14</sup>

Still, there are some large schools in urban areas, where student sorting is possible. Since experienced teachers are more often in urban areas, this may explain why selection on initial scores can be observed for classes with experienced teachers and not for novice teachers.

Finally, we will estimate the class size effect, again using the sample that excludes experienced teachers. As the correlation between class size and observed initial scores is significant on the whole sample, we suspect that there may also be a selection on unobserved variables, which could perturb the estimation of the causal effect of class size on pupils' achievement. On the contrary, the correlation between class size and initial scores is not significant for the sub-sample of pupils with novice teachers. Hence we assume that the "traditional" selection bias is taken care of. Finally, to assess the

robustness of our findings, we will also estimate the class size effect on all the students, using instrumental variables.

### 3. Estimation Results of the Training Effect

Estimation results of the effects of teacher and class characteristics on pupils achievement are presented in Table 4. It is a regression of final test scores on initial (and national) test scores, student, teacher, and class characteristics. The estimation controls for class effects and is estimated on the sub-sample of students with novice teachers.

Table 4

*Estimated Effects of Teacher and Class Characteristics on Final Test Scores*

*Notes.* Regression of the **final** test scores of **the students with novice teachers**; when the covariates are included, they comprise student and class characteristics (including initial test scores), together with teacher characteristics. Coefficients are estimated through mixed models, with class effects. Standard errors are reported in parentheses. \*\* $p < 0.05$  \* $p < 0.10$

The regressions of final test scores include class characteristics. Some class characteristics can be computed using the individual characteristics. In particular, we computed the fraction in the class of students with advantaged parents as measured by occupations, the fraction of girls, the fraction of non-French students, and the fraction of students who repeated at least one grade. These variables are calculated for each student, excluding his/her own characteristics in the calculation of the means. None of these variables yields significant coefficients. They are not included in the specification presented in this paper.

On the contrary, means and standard deviations of initial test scores within classes have significant correlations with final scores: both means and standard deviations of initial test scores have a negative impact on pupils' improvement, meaning that students have better results in homogeneous classes and when the average achievement is not too high.<sup>15</sup>

The estimated impact of teachers' training is not significant for reading achievement but it is significant and large for mathematics achievement: students gain one quarter of a standard deviation on their final scores when their teachers have been trained. This effect should be compared to the coefficient estimated on initial test score (0.062). The effect on final scores is substantial enough to be confident that it is not driven by a bias in the assignment of teachers to classes.

These results are close to the findings reported in Bressoux (1996). They are also close to the raw differences of the means: as can be seen from Table 1, students with untrained novice teachers have similar initial scores than students with trained novice teachers; yet, their performance is much lower during the year. The raw difference-in-difference estimator gives an effect of 0.14 in reading, and 0.25 in mathematics. Incorporating other covariates decreases the effect in reading, close now to 0.05 and not significant. The estimation of the regression without class effects would have led to a significant effect equal to 0.11. Hence incorporating (correlated) random class effects makes the training coefficient weaker and non significant.

It would be very interesting to fully understand what is efficient in the training: is it the improvement of pure academic skills in mathematics or the improvement in

educational and teaching methodologies (including on-the-job experience even if the training in France does not include much of this kind of training)?

Our data are not precise enough to thoroughly explore this question. Nevertheless they include a lot of information about the teachers' instructional practices. They include in particular the number of hours per week used for teaching reading or mathematics, the number of hours asked for homework per week, the practice of organizing the class in groups, and how these groups are chosen. Yet none of these variables is significantly correlated to the dummy for trained/untrained teachers.<sup>16</sup> So it seems that trained and untrained teachers have similar practices. Furthermore, these variables have been tested in the regression of the final test scores on the initial test scores, student and class characteristics. None of the coefficients are significant, meaning that none of the practices seem to have better results than others. The small number of classes in the sample may prevent us from identifying effects that might exist.

The data also give some information about the teachers' academic backgrounds, in particular the diploma and the discipline studied at university. So we can add this information in our regression. More precisely, we can include indicator variables for teachers having majored in sciences at university (14% of novice teachers) and for teachers having majored in a discipline not reported in the survey (roughly 14% of novice teachers). This latter group of teachers comprises some individuals who did not report this information, potentially because of multiple fields of specialization, as well as a small number of teachers who did not go to university. Novice teachers are endowed with similar diplomas since it is compulsory to have a degree equivalent to two years in a university to enter an 'école normale'. Yet parents of three or more children and professional sports-persons are exempt from the diploma requirement. The last category finally comprises those teachers who majored in humanities (often French or another language, sociology, psychology, history).

Our estimates including the teachers' academic background show that, even though the training effect is substantial in mathematics achievements, teachers who have not been trained, but who have studied mathematics or sciences at university, compensate for this lack of training. Training seems to be very useful in improving scientific knowledge for teachers who studied other disciplines in their university. Yet, caution is granted because the sample is small.

Thanks to the structure of our data, comprising the information about all the students of a given class, we can estimate heterogeneous training effects on students, within each class. These effects are measured by interacting the indicator variable for trained teachers with quartiles constructed using the distribution of students scores within a class.

The quartiles are constructed using students' initial test scores in reading and are measured within the class. Thanks to these quartiles, we have four categories of students: low-achieving students, medium low-achieving students, medium high-achieving students, and high-achieving students. In the estimation, we exclude the training variable (meaning the dummy for having a trained teacher) and we include the four interactions between the training variable and the student categories. This is done in a unique regression, to keep controlling for class effects. This methodology allows us to estimate the within-class effect for these student-specific variables.

No significant training effects appear for scores in reading (Table 5). For scores in mathematics, the training effect is substantial and significant for all students except low-

achieving students. Training helps teachers improve students' results in mathematics, except for the least able ones.

Table 5  
*Differential Effects of Training*

*Notes.* Regression of the **final** test scores of **the students with novice teachers**; the covariates include student, class characteristics and dummy variable of trained teachers, broken down according to the quartiles. Coefficients are estimated through mixed models, with class effects. Standard errors are reported in parentheses. \*\*p<0.05 \*p<0.10

Finally, to test for the robustness of the training effect, we estimated a propensity score for a class to have a trained teacher using a logistic model at the class level. Class means of initial test scores do not differentiate trained and untrained teachers, except the standard deviation of initial math scores. On the contrary, trained teachers have a much lower probability to be in a rural or semi-rural school than untrained teachers. To discard extreme classes which could bias our estimates, we kept only classes which are within the common support of both distributions. The results (available from the authors) confirm the preceding ones: the correlation between the trained teacher indicator and initial scores is even smaller than before, and the training effects estimated on final scores are very similar to those estimated on the whole sample of novice teachers.

#### **4. Estimation Results on the Class Size Effect**

Results for the class size effect are also contained in Table 4. Class size has a significant impact on students' outcomes. The impact is quite similar in reading and in mathematics, close to 3% of a standard deviation of final test scores. This result should be compared to class size correlations with initial test scores (Table 3): these correlations are not negligible, around 1% but they are *positive*, meaning that the effects measured on final test scores could even be biased downwards.

This impact seems to be robust to the problem of combination classes. The regression is estimated with a sample including multiple-grade classes. In the case of a combination class, the class size is then the size of the entire class, and not the number of third-graders. Yet the indicator for multiple-grade classes is not significant.<sup>17</sup> Results are similar when excluding these combination classes.

This impact is substantial: reducing the class size by 10 students increases the final test scores by 25 to 30 percentage points. Using similar arguments as presented for teachers' training, we believe our results come from a quasi-experimental design and are, therefore, causal. This estimated impact is close to that obtained for teachers' training in mathematics. Comparing to other studies on class size, it is close to the highest estimates, in particular it is roughly equal to the effect reported in the Tennessee STAR experiment (Krueger, 1999).

One strand of research about class size focuses on the channels which could explain how a small class may help students improving their scores. In our data, the information given by the teacher about his or her practices in the class can be used to partially open the black box and to evaluate whether training effect or class size effect are driven by different teaching practices.

The regressions are done at the class level. We estimate whether class size can be explained by the variables describing the teaching practices.<sup>18</sup> None of the practices seems correlated with class size. The same is true with a dummy variable for large classes, instead of the class size variable. This result is consistent with other studies, which usually find that teaching practices do not vary with class size (for a survey, see Ehrenberg, Brewer, Gamoran and Willms, 2001).

Some papers find that a reduction in class size affects more directly low-achieving students. However, this result may come from some composition effect, hiding the fact that class size benefits similarly to all students within a class. We use the same decomposition in quartiles within classes as for the training effect.

The results for differential effects of class size are given in Table 6. Within classes, the pattern for scores in reading is relatively flat, the class size affecting similarly all students, except a slightly smaller effect for low-achieving students. On the contrary, the effect on mathematics scores decreases when the “quality” of the students increases. Hence, within classes, low-achieving students seem to benefit more from being in a small class than high-achieving students.

Table 6  
*Differential Effects of Class Size*

*Notes.* Regression of the **final** test scores of **the students with novice teachers**; the covariates include student, class characteristics and class size variable, broken down according to the quartiles. Coefficients are estimated through mixed models, with class effects. Standard errors are reported in parentheses. \*\*p<0.05 \*p<0.10

Heterogeneity in the class size effect is confirmed by results when estimating the class size effect for all schools and for ZEP schools alone, estimated in the same regression. The class size effects are larger in the ZEP school, and are very substantial: 5% in reading (yet non-significantly different from the main effect of 2%) and 8% in mathematics. This finding confirms recent results obtained by Piketty (2004) who also finds a substantial impact of class size in ZEP schools, albeit marginally significant because of the small number of students in ZEP schools in his sample.

These results show that the students in disadvantaged schools are more sensitive to class size as a group than other groups of students. One potential explanation is behavioural; in larger classes the probability of having a troublemaker among students is larger in these schools (see Lazear, 2001).

The effect of third-grade class size, as estimated in this paper, stands between 2.5 and 3 percentage point of final test scores. Piketty (2004), on second-grade classes, finds similar effects. He applies a methodology developed by Angrist and Lavy (1999). His method is based on the following specificity of French class openings, similar in that to Israel’s case studied in Angrist and Lavy: when second-grade enrolment goes beyond 30, another class is opened (in most cases). Hence, the two new classes have an average size of 15 pupils. Piketty uses this discontinuity as an instrumental variable. He finds that a reduction in class size induces a significant and substantial increase in mathematics and reading scores, and that the effect is larger for low-achieving students.

In our data, we find similar specificities as those observed by Piketty (2004).<sup>19</sup> There are often two classes when the number of third-graders in the school is larger than 30.

Yet, some classes have up to 34 students. And the link between class size and enrolment is loose because of the possibility of combination classes. When the enrolment goes beyond 30 students, the schools do not automatically open another third-grade class, but may instead assign some third-graders to classes with students of other grades.

Nevertheless, to check the robustness of our class size effects estimated for the novice teachers, we use instrumental variables on the whole sample. The instrument is based upon the enrolment of third-graders in the school when we exclude combination classes. Another strategy is to work with “pure” third-grade classes and combination classes, including in the enrolment the students who are in the same class as some third-graders, whatever grade they are in (see Figure 2). In all cases, the instrumental variable is the mean of the class sizes in the school: the sizes of the “pure” third-grade classes in the first case and those of all classes with third-graders in the second case. This instrument takes care of the selection bias which arises when schools organize classes so that small classes gather low-achieving students and high-achieving students are assigned to larger classes.

Fig. 2  
*Enrolment and Size of Third-Grade Classes*

As can be seen on Figure 2, the instrumental variable is very close to the actual class size. Indeed, in our data, we identify few schools with more than one third-grade class. And when there are two classes, the sizes of these two classes are not very different. Hence the findings are easy to foresee: the results estimated with the instrumental variable are very close to the OLS results (see Table 7).

Table 7  
*Effect of Class Size Estimated with or without Instrumental Variable*

*Notes.* Regression of the **final** test scores of **all the students**; the covariates include student, class characteristics and instrumented variable of class size. The last column correspond to the estimation on the sub-samples restricted to enrollments between 24 and 45 students. Coefficients are estimated through mixed models, with class effects. Standard errors are reported in parentheses. \*\*p<0.05 \*p<0.10

The idea in Angrist and Lavy (1999) is to use the discontinuity of the class size resulting from the creation of several classes when the enrolment goes beyond some level, assuming that this discontinuity is exogenous. One way of using this discontinuity is to estimate the class effect only when the enrolment is close to the “breaking point”. We also use this method and estimate our instrumented regression for school where the enrolment is close to 34 students, the “breaking point” according to our data. We have chosen to restrict the sample to enrolments between 29 and 40. The coefficients are still substantial, even if they are no more significant.

All these results confirm the estimated magnitude: the class size effect is close to 2.5 percentage point of the final test scores.

## 5. Conclusion

This paper looks at the impact of teachers’ training using a quasi-experimental design together with the administrative mistakes made in forecasting the required number of

teachers, and checking using observable variables that the assignment of teachers is as good as random. The effect of teachers' training is substantial: final test scores in mathematics of students with a trained teacher are larger by 25% of the standard deviation than the scores they would have had if their teachers had not been trained.

This estimation should not confound the effect of training with the effect of age since trained and untrained teachers are on average the same age. Nevertheless, this training effect could reveal an experience effect, through the on-the-job experience included in the training program. Due to the small sample, we cannot be completely conclusive on what is behind this training effect: improvement of academic skills, improvement in educational methodologies or on-the-job experience. Yet, using the educational background, we find some evidence that training is efficient in improving discipline competence in mathematics, in particular for those teachers who did not study sciences at their university. This finding confirms other studies which find an effect of the academic background of the teachers.

The effect of class size is shown to be significant and negative: a smaller class size improves student achievement. The impact is evaluated between 2.5% and 3% of a standard deviation of the scores. Hence, training teachers is equivalent to reducing class size by 10 students, in terms of final test scores in mathematics. It is worth noting that the effect of class size seems more beneficial to low-achieving students within classes. The effect is particularly large for classes in priority education areas. On the contrary, it seems that low-achieving students do not benefit from the training of their teachers. This finding shows the complexity of the education production function and proves that it is essential to study how resources impact different students differently.

## References

- Altonji, J. G., Elder, T. E. and Taber, C. R. (2005), 'Selection on observed and unobserved variables: assessing the effectiveness of catholic schools', *Journal of Political Economy*, vol. 113(1) (February), pp. 151-184.
- Angrist, J. D. and Lavy, V. (1999), 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement', *Quarterly Journal of Economics*, vol. 114(2) (May), pp. 533-574.
- Angrist, J. D. and Lavy, V. (2001), 'Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools', *Journal of Labor Economics*, vol. 19(2) (April), pp. 343-369.
- Benabou, R., Kramarz, F. and Prost, C. (forthcoming), 'The French Zones d'Éducation Prioritaire: much ado about nothing', *Economics of Education Review*.
- Boyd, D, Grossman, P., Lankford, H., Loed, S. and Wyckoff, J. (2006), 'How changes in entry requirements alter the teacher workforce and affect student achievement', *Education Finance and Policy*, vol. 1(2) (March), pp. 176-216.
- Bressoux, P. (1996), 'The effects of teachers' training on pupils' achievement: the case of elementary schools in France', *School Effectiveness and School Improvement*, vol. 7(3) (Fall), pp. 252-279.
- Brizard, A. (1995), 'Écoles rurales, écoles urbaines : performance des élèves en français et en mathématiques', *Éducation et Formations*, vol. 43 (Octobre), pp. 105-111.
- Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2006), 'Teacher-student matching and the assessment of teacher effectiveness', *Journal of Human Resources*, vol. 41(4) (Fall), pp. 778-820.
- Ehrenberg, R. G. and Brewer, D. J. (1994), 'Do school and teacher characteristics matter? Evidence from *High School and Beyond*', *Economics of Education Review*, vol. 13(1) (March), pp. 1-17.
- Ehrenberg, R. G., and Brewer, D. J. (1995), 'Did teachers' verbal ability and race matter in the 1960s? *Coleman revisited*', *Economics of Education Review*, vol. 14(1) (March), pp. 1-21.
- Ehrenberg, R. G., Brewer, D. J., Gamoran, A. and Willms, D. J. (2001), 'Class size and student achievement', *Psychological Science in the Public Interest*, vol. 2(1) (May), pp. 1-30.
- Hanushek, E. A. (2003), 'The failure of input-based schooling policies', *ECONOMIC JOURNAL*, vol.



- 113(485) (February), pp. 64-98.
- Hoxby, C. M. (2000), 'The effects of class size on student achievement: new evidence from population variation', *Quarterly Journal of Economics*, vol. 115(4) (November), pp. 1239-1285.
- Kane, T. J., Rockoff, J. E. and Staiger, D. O. (forthcoming), 'What does certification tell us about teacher effectiveness? Evidence from New York City', *Economics of Education Review*.
- Krueger, A. B. (1999), 'Experimental estimates of educational production functions', *Quarterly Journal of Economics*, vol. 114(2) (May), pp. 497-532.
- Krueger, A. B. (2003), 'Economic considerations and class size', *ECONOMIC JOURNAL*, vol. 113(485) (February), pp. 34-63.
- Lazear, E. P. (2001), 'Educational production', *Quarterly Journal of Economics*, vol. 116(3) (August), pp. 777-803.
- Moulton, B. R. (1986), 'Random group effects and the precision of regression estimates', *Journal of Econometrics*, vol. 32(3) (August), pp. 385-397.
- Oeuvarard, F. (1995), 'Les performances en français et en mathématiques des écoles à classe unique', *Éducation et Formations*, vol. 43 (Octobre), pp. 113-116.
- Piketty, T. (2004), 'L'impact de la taille des classes et de la ségrégation sociale sur la réussite scolaire dans les écoles françaises : une estimation à partir du panel primaire 1997', mimeo.
- Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005), 'Teachers, schools, and academic achievement', *Econometrica*, vol. 73(2) (March), pp. 417-458.
- Robinson, G. K. (1991), 'That BLUP is a good thing: the estimation of random effects', *Statistical Science*, vol. 6(1) (February), pp. 15-32.
- Rockoff, J. (2004), 'The impact of individual teachers on student achievement: evidence from panel data', *American Economic Review*, vol. 94(2) (May), pp. 247-252.
- Summers, A. and Wolfe, B. (1977), 'Do schools make a difference?', *American Economic Review*, vol. 67(4) (September), pp. 639-652.
- Thaurel-Richard, M. (1995), 'Les progrès des élèves au CE2 en milieu rural', *Éducation et Formations*, vol. 43 (October), pp. 117-123.
- Woessmann, L. (2005), 'Educational production in Europe', *Economic Policy*, vol. 20(43) (July), pp. 445-504.
- Woessmann, L. and West, M. R. (2006), 'Class-size effects in school systems around the world: evidence from between-grade variation in TIMSS', *European Economic Review*, vol. 50(3) (July), pp. 695-736.

#### Footnotes:

1. See, amongst others, Rivkin, Hanushek and Kain (2005) and Rockoff (2004).
2. See for instance Krueger (1999). The estimation of the class size effect continues to yield a huge literature. For some recent papers on the topic, see among many others Angrist and Lavy (1999), Ehrenberg, Brewer, Gamoran and Willms (2001), Hanushek (2003), Hoxby (2000), Krueger (2003), Woessman and West (2006).
3. Since 1992, the requirement to take the exam is a three-year-university diploma.
4. France is geographically divided into administrative 'départements' and there was an 'école normale' in each 'département'. After 1991, the 'écoles normales' were closed and since then, the new teacher training colleges were called university teacher training institutes 'Instituts Universitaires de Formation des Maitres (IUFM)' and started to depend administratively on a region (a region includes several departments). The new training period (after the exam) became one year.
5. These errors are likely to have been amplified by future changes in the teachers' training institutions: from Ecoles Normales to IUFM and from two years of training to one year. These changes happened just after the specific year of our study.
6. These statistics are slightly different from the ones in Bressoux (1996) because of some small differences in the methods for matching the student data and the teacher data.
7. In all the regressions, the information about parents' occupations is summarized through a dummy variable depicting high occupations, corresponding to self-employed worker, executive, teacher, professor, technician, and foreman.

8. All our estimates are very similar without these imputations.
9. Priority educational areas are enrolled in a specific program which will be briefly described below.
10. We have not implemented tests such as the one described in Altonji, Elder, and Taber (2005) because of the relatively small number of teachers in our data, and also because we are relatively confident that a selection on unobservables would be accompanied with a selection on initial test scores.
11. The priority zones are more often in urban areas, where classes are larger than in rural areas. So the effective reduction in class size in ZEP schools could be larger than the one given by the raw difference of the two means. A regression of the class size on the dummy variable for ZEP schools, controlling for the rural areas and the combination classes, shows that class size is smaller by 1.75 students in priority zones.
12. We obtain these results with our data. They are consistent with Brizard (1995) and Thuarel-Richard (1995).
13. In the ZEP junior high schools, there is some evidence that enrolment decreases because these areas are avoided by some parents (see Benabou, Kramarz, Prost, forthcoming). To counter this flight, some principals give better conditions to the most affluent students. Yet we do not observe this effect in the primary schools of our data. On the contrary, as we will see later, the positive correlation between class size and initial achievement is even larger in ZEP schools.
14. In our data, 77% of the schools have less than 11 classrooms, for 5 grades (1<sup>st</sup> grade to 5<sup>th</sup> grade). But urban areas are overrepresented in our sample. In 2006, in France, 62% of the schools have less than 6 classrooms, and 92% less than 11 classrooms.
15. This is a raw result which should not lead to any conclusion. Exploring the question of the impact of class homogeneity is not the purpose of this paper.
16. Results are available from the authors.
17. This result is consistent with Oeuvarard (1995).
18. Results are available from the authors.
19. In contrast with Piketty (2004), we do not always observe all third-grade classes in schools.

Figure 1

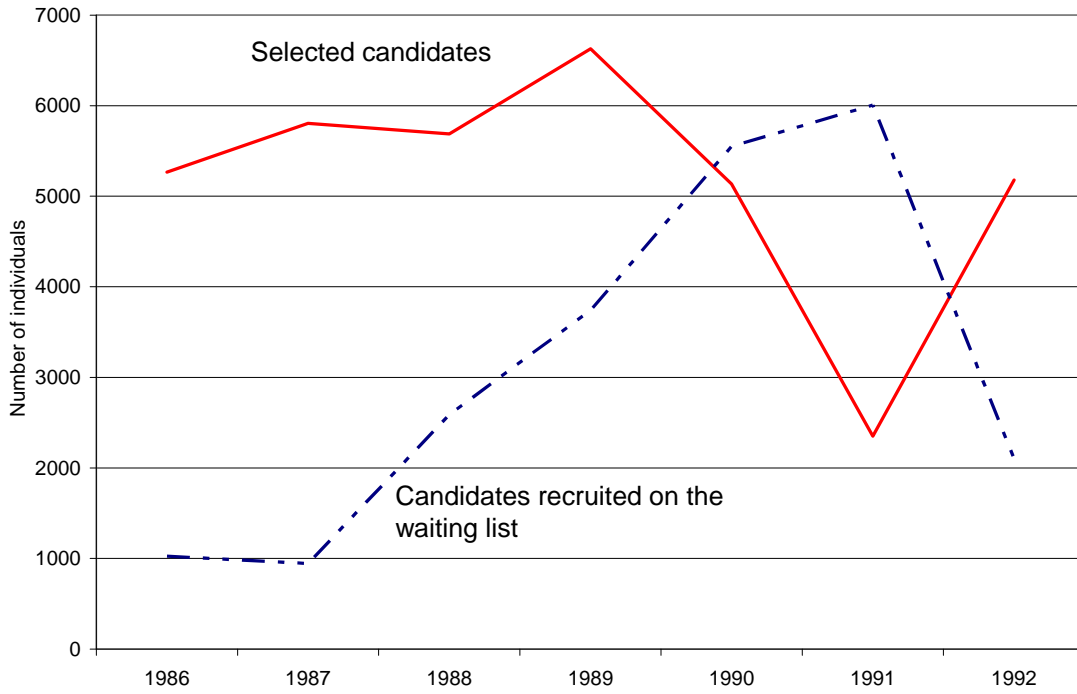
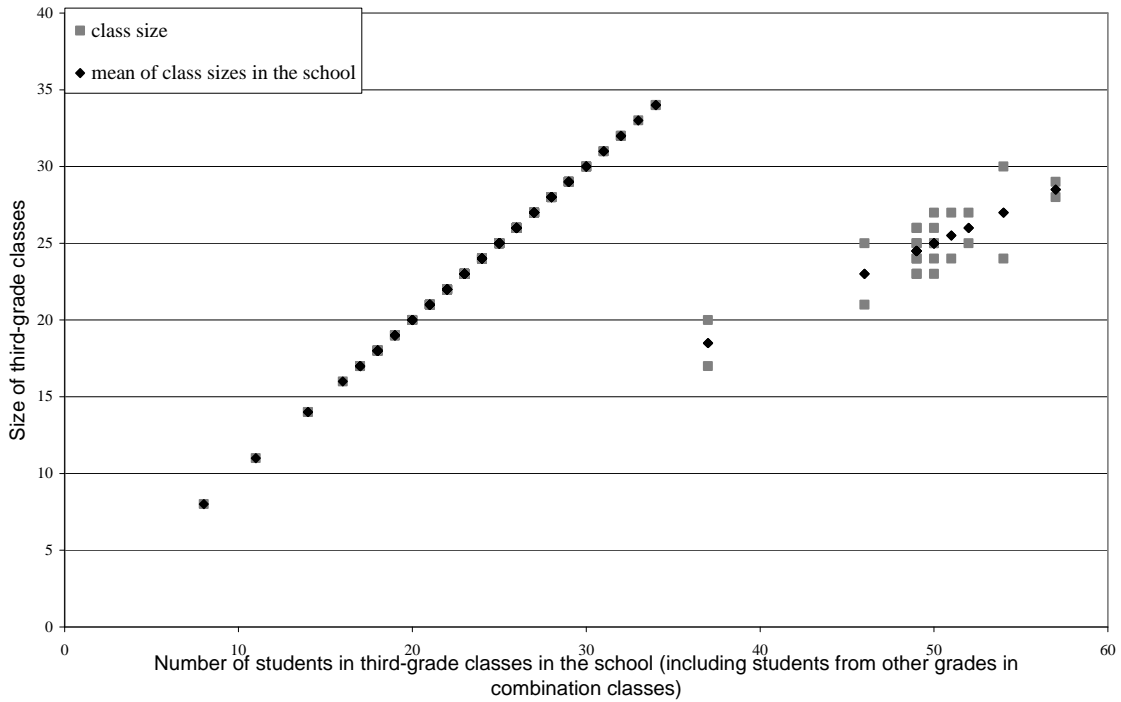


Figure 2



**Table 1**

	Experienced teachers	Trained novice teachers	Untrained novice teachers	Chi-squared tests across the 3 categories	Chi-squared tests trained versus untrained
Variable	Mean	Mean	Mean	P-value	P-value
<i>A. Student variables</i>					
Month of birth	6.51	6.55	6.61	0.69	0.46
One year younger than usual age	0.02	0.01	0.02	0.79	0.57
Female	0.48	0.48	0.47	0.98	0.83
Foreign nationality	0.14	0.21	0.17	0.00	0.12
1 sibling	0.40	0.29	0.35	0.00	0.01
2 siblings	0.26	0.28	0.29	0.19	0.59
3 or more siblings	0.25	0.35	0.25	0.00	0.00
Father's occupation: high	0.35	0.29	0.33	0.01	0.19
Mother's occupation: high	0.18	0.15	0.14	0.01	0.90
Pre-elementary school: less than 3 years	0.13	0.15	0.16	0.08	0.41
Pre-elementary school: more than 3 years	0.20	0.17	0.16	0.06	0.88
One repeated grade in pre-elementary school	0.01	0.02	0.01	0.27	0.38
First grade repeated	0.09	0.15	0.14	0.00	0.49
Second grade repeated	0.07	0.07	0.05	0.29	0.12
Third grade repeated	0.07	0.06	0.05	0.17	0.27
Initial test score in reading	5.15 (0.95)	4.79 (1.01)	4.85 (1.00)		
Initial test score in math	5.15 (0.98)	4.82 (0.99)	4.84 (0.98)		
Final test score in reading	5.16 (0.96)	4.81 (1.01)	4.73 (1.00)		
Final test score in math	5.14 (0.96)	4.91 (0.99)	4.68 (1.04)		
<i>Number of students</i>	1695	1027	487		
<i>B. Teacher and class variables</i>					
Female	0.59	0.81	0.88	0.00	0.42
Age	43.46 (6.51)	28.13 (6.11)	27.84 (6.93)		
Father's occupation: high	0.26	0.49	0.53	0.00	0.72
Mother's occupation: high	0.09	0.19	0.16	0.17	0.68
Field of specialization: sciences	0.10	0.14	0.13	0.70	0.81
Field of specialization: unknown	0.70	0.13	0.19	0.00	0.43
Class size	24.95 (3.65)	22.73 (4.05)	23.53 (4.85)		
Large class (>25 students)	0.41	0.21	0.34	0.03	0.15
Combination class	0.17	0.37	0.44	0.00	0.49
Rural	0.12	0.11	0.28	0.06	0.04
Semi-rural	0.15	0.19	0.28	0.25	0.31
Priority educational area (ZEP)	0.20	0.35	0.28	0.11	0.50
<i>Number of classes</i>	82	63	32		

**Table 2**

Dependent variable:	Initial scores in reading			Initial scores in math		
Experienced teacher	0.318 ** (0.116)		0.269 ** (0.094)	0.294 ** (0.112)		0.264 ** (0.096)
Trained novice teacher	-0.065 (0.123)		0.052 (0.098)	-0.082 (0.118)		0.038 (0.099)
Class size		0.025 ** (0.011)	0.016 * (0.009)		0.019 * (0.01)	0.010 (0.009)
Other variables	No	No	Yes	No	No	Yes
Class effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of students	3718	3650	3650	3584	3521	3521
Number of classes	196	188	188	188	181	181
Class level residual variance	0.289	0.308	0.161	0.256	0.278	0.157
Individual level residual variance	0.701	0.698	0.561	0.727	0.726	0.633

**Table 3**

Dependent variable:	Initial scores in reading			Initial scores in math		
Trained novice teacher	-0.065 (0.130)		0.068 (0.111)	-0.082 (0.118)		0.062 (0.101)
Class size		0.015 (0.015)	0.014 (0.013)		0.009 (0.014)	0.008 (0.012)
Other variables	No	No	Yes	No	No	Yes
Class effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of students	1711	1711	1711	1685	1685	1685
Number of classes	100	100	100	98	98	98
Class level residual variance	0.327	0.325	0.201	0.255	0.255	0.152
Individual level residual variance	0.721	0.721	0.618	0.727	0.727	0.666

**Table 4**

Dependent variable:	Final scores in reading			Final scores in math		
Trained novice teacher	-0.024 (0.120)		0.048 (0.077)	0.170 (0.122)		0.241 ** (0.091)
Class size		-0.021 (0.014)	-0.026 ** (0.009)		-0.029 * (0.014)	-0.027 ** (0.010)
Other variables	No	No	Yes	No	No	Yes
Class effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of students	1680	1680	1605	1671	1671	1595
Number of classes	100	100	98	98	98	96
Class level residual variance	0.266	0.258	0.095	0.274	0.267	0.128
Individual level residual variance	0.761	0.761	0.251	0.789	0.789	0.318

**Table 5**

Dependent variable:	Final scores in reading	Final scores in math
Trained teacher and low achieving students (Q1)	0.053 (0.085)	0.119 (0.097)
Trained teacher and medium low achieving students (Q2)	0.023 (0.082)	0.269 ** (0.097)
Trained teacher and medium high achieving students (Q3)	0.022 (0.082)	0.271 ** (0.097)
Trained teacher and high achieving students (Q4)	0.098 (0.086)	0.334 ** (0.099)
Other variables	Yes	Yes
Class effects	Yes	Yes
Number of students	1605	1569
Number of classes	98	96
Class level residual variance	0.095	0.129
Individual level residual variance	0.251	0.316

**Table 6**

Dependent variable:	Final scores in reading	Final scores in math
Class size and low achieving students (Q1)	-0.022 ** (0.009)	-0.035 ** (0.010)
Class size and medium low achieving students (Q2)	-0.026 ** (0.009)	-0.028 ** (0.010)
Class size and medium high achieving students (Q3)	-0.027 ** (0.009)	-0.027 ** (0.010)
Class size and high achieving students (Q4)	-0.027 ** (0.009)	-0.024 ** (0.010)
Other variables	Yes	Yes
Class effects	Yes	Yes
Number of students	1605	1569
Number of classes	98	96
Class level residual variance	0.095	0.130
Individual level residual variance	0.250	0.313

**Table 7**

Dependent variable:	All classes						Without combination classes					
	Final scores in reading			Final scores in math			Final scores in reading			Final scores in math		
	OLS	IV	IV (24-45)	OLS	IV	IV (24-45)	OLS	IV	IV (24-45)	OLS	IV	IV (24-45)
Class size	-0.024 ** (0.006)	-0.025 ** (0.007)	-0.019 (0.014)	-0.025 ** (0.007)	-0.025 ** (0.007)	-0.017 (0.014)	-0.023 ** (0.007)	-0.023 ** (0.007)	-0.012 (0.014)	-0.026 ** (0.008)	-0.026 ** (0.008)	-0.016 (0.012)
Other variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Class effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of students	3447	3447	2248	3347	3347	2129	2924	2924	1971	2826	2826	1844
Number of classes	185	185	106	179	179	100	132	132	82	127	127	76