International Journal of Educational Development 30 (2010) 544–559

Contents lists available at ScienceDirect

International Journal of Educational Development

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

Does taking one step back get you two steps forward? Grade retention and school performance in poor areas in rural China

Xinxin Chen [a,*], Chengfang Liu [b], Linxiu Zhang [b], Yaojiang Shi [c], Scott Rozelle [d]

[a] College of Economics, Zhejiang Gongshang University, 18 Xuezheng Street, Xiasha University Town, Hangzhou 310018, China
[b] Center for Chinese Agricultural Policy, Institute of Geographical Sciences and Natural Resource Research, Chinese Academy of Sciences, No. 11A Datun Road, Anwai, Beijing 100101, China
[c] Northwest Social Science Development Research Center, Northwest University, Xi'an 710067, China
[d] Freeman Spogli Institute, Stanford University, 616 Serra Street, CA 94305, USA

ARTICLE INFO

JEL classification:
I21
I28
O53

Keywords:
Grade retention
School performance
Human capital

ABSTRACT

Despite the rise in grade retention in poor areas in rural China recently, little work has been done to understand the impact of grade retention on the educational performance of students in these areas in rural China. This paper seeks to redress this shortcoming and examines the effect of grade retention on educational performance on 1649 students in 36 elementary schools in Shaanxi province. With a dataset that was collected from a survey designed specifically to capture school performance of students before and after they were retained, we use differences-in-differences, propensity score matching and differences-in-differences matching approaches to analyze the effect of grade retention on school performance. Although the descriptive analysis shows that grade retention helps to improve the scores of the students that were retained, somewhat surprisingly, the results from the multivariate analysis consistently show that there is no significant positive effect of grade retention on school performance of the students. In fact, in some cases (e.g., for the students who repeat grade 2), grade retention is shown to hurt school performance.

1. Introduction

In the past, because of fiscal considerations, many provincial Departments of Education in China (and other developing countries) abolished retention or set maximum limits on the percentage of students in each cohort who could be retained during their primary school years (Guangming Daily, 2000; Liddell and Rae, 2001). For example, in Gansu province in China, primary school principals were instructed to retain no more than 3% of the students in any of the six grades that constitute elementary education (Guangming Daily, 2000). The reasoning was that since China was relatively poor, fiscally, mass education demanded that students move rapidly through the periods of compulsory education.

While the logic of such policies may be understandable from a budgetary perspective, the experience of educators internationally suggests that such a policy might adversely affect a certain segment of students. Specifically, the empirical literature outside of China has shown that in some cases students benefit from grade retention (Alexander et al., 1994; Kerzner, 1982). The argument is that grade retention is good for students who are behind in their studies (and receiving failing or near failing grades) since they are allowed to relearn the material and catch up with their peers. If this literature is correct, those who are retained should show relative improvement in the years after they were retained.

In recent years, perhaps because China's fiscal situation has improved dramatically (Wang and Zhou, 2005), the Ministry of Education eliminated the restrictions on the maximum number of students that could be retained (Ministry of Education, 2006). In some provinces local school administrators have been granted more authority in deciding how many students would be retained each year. In addition, in some areas the funding formulas were relaxed to allow schools who had relatively high rates of retention to continue to receive a "per student" subsidy (or allocation) for all students, even for those students that were repeating grades (CCTV, 2006). The idea was that this might improve the quality of education, a result that would be consistent with the findings cited above. However, it has been reported that in some localities school officials may have taken advantage of the new compensation rules and artificially inflated the retention rates purely for the fiscal gain for the school (CCTV, 2006).

* Corresponding author. Tel.: +86 571 88211232; fax: +86 571 28008035.
E-mail address: xinxchen@yahoo.com (X. Chen).

0738-0593/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijedudev.2009.12.002

As retention rates rose in some areas, new concerns, also grounded in the debate within the international education literature, surfaced. In contrast to the empirical literature that finds retention improves school performance of students (as discussed above), other research criticizes the casual use of grade retention (Grissom and Shepard, 1989; Holmes, 1989; Fine, 1991). While students who repeat a grade do get a chance to catch up, they also may experience negative psychological effects. Some educators believe that grade retention destroys the self-respect and confidence of students and can actually decrease educational performance (Royce et al., 1983). There also is a cost to the family, which has to pay for the associated costs of another year of education, and retention extends the time that their child is in school, delaying his/her entrance into the labor force (Yang, 1991). If grade retention is associated with poorer education performance, then local policies that encourage high rates of grade retention could systematically be hurting students.

Somewhat surprisingly, little work has been done to understand whether or not grade retention in the context of China helps or hurts the educational performance of children. There are discussions of grade retention in China's social science literature (Huang, 1998; Yu, 1999; Wen, 2002; Li, 2004). But, these papers are at best based on descriptive statistics. Most of the work is based on case studies and uses anecdotes as evidence. Given the fundamental importance in trying to develop better policies for improving education, there is a need to more rigorously understand empirically how grade retention affects school performance.

The overall goal of this paper is to examine the effect of grade retention on the educational performance of elementary school students in poor areas in rural China. It is possible that a better understanding of the impact of grade retention will provide policy makers with the information they need to make (or not make) changes to the administration of the educational system in China's rural areas. The results can also contribute to the general understanding of the relationship between grade retention and school performance. To meet this goal, we pursue three specific objectives. First, we compare the change of scores over time of students that were retained with students that were not retained. Second, we examine the determinants of grade retention in rural China in order to find what types of students are most likely to repeat a grade during their elementary school years. Third, we examine with multivariate analysis whether or not grade retention improves or hurts the school performance of rural students by comparing their performance relative to their fellow students, both before and after they were retained.

To meet these objectives, we will rely on a set of data that we collected in 2006, a data collection effort that was designed specifically to examine changes in school performance of children before and after they were retained. With this data set, we focus our attention on two types of students: the students who were retained in grade 2, grade 3 or grade 4 and their fellow students in the same grades that were not retained. Using these different subsets of students, we compare changes in scores before and after the students repeated their grades. A descriptive analysis is supplemented by a more rigorous multivariate analysis on the effects of grade retention on the educational performance using several approaches, including a differences-in-differences approach (DID), propensity score matching (PSM) and a combination of these two approaches (DIDM).

This study is unique in several respects. First, it contributes to the limited understanding of the effects of grade retention on the school performance in China by examining how children's scores correlate with grade retention. To date the empirical literature on grade retention and school performance in China is almost nonexistent. Second, we use the most up-to-date evaluation methods, instead of the more traditional descriptive/case study/Ordinary Least Squares approaches.

There are limitations in our approach, however. For example, we focus on students from one province in China's northwest region, Shaanxi province. Since Shaanxi province is in one of the poorest parts of China, this limits our ability to say anything about China in general. In addition, since we only examine the effect of grade retention on the school performance of the students who were retained in grade 2, grade 3 or grade 4, our conclusions can not necessarily be generalized to those students who are retained in grade 1 (the most common grade during which students are retained) or in any grade that is greater than grade 5.

2. Data

The data used in this paper come from a survey executed by the authors in 2006. The survey was designed specifically to examine the changes in school achievement of children before and after they repeated at least one grade. While the survey in part relied on recall data—especially for some of the control variables—we were able to use records and rely on multiple sources of information for our two key variables—scores of school achievement and grade retention.

The sample was drawn from 36 primary schools in 12 towns in Shaanxi province, one of the nation's poorest provinces in northwest China. The sample was drawn using a multi-stage, clustering design with random selection procedures employed at each stage. In the first stage six counties were selected from the total of 93 counties in Shaanxi province. In the second stage the survey team randomly selected two townships in each county. The two townships were chosen from a list of all townships in the county that were ranked according to per capita income. One township was chosen from the townships that were relatively rich and the other from the townships that were relatively poor. In the third stage a list of all primary schools was created in each township (where schools were limited to all primary schools that included six years of schooling—or all *wanxiao*). From this list three primary schools were chosen randomly.

The sample students were selected during the final stage of the sampling. The sample design consisted of all students that were in the sixth grade classes in each of the sample schools when they were interviewed.[1] On average there were 1.4 sixth grade classes per school, ranging from one to three. When the data were collected in September, the students had just begun a new school year. Therefore, all of the sample students had just completed the fifth grade less than two months previously (as the school year in China runs between early September and mid-July). In total, the sample included 1653 children and their families. Approximately 45% of sample students were girls. The ages of the students ranged between 10 and 16; however, most of the students (73%) were either 11 or 12.

Our main measure of education achievement is based on the Chinese language scores of the students from 2001/2 (their first grade year) to 2005/6 (their fifth grade year). Fortunately, in China every student in almost every elementary school (including, at least, all of the schools in our sample) keeps in his/her possession a booklet that contains a comprehensive record of the Chinese language scores for each semester of his/her schooling. This means that the school achievement variables that we use in our analysis

[1] We collected information on the number of students in the sixth grade in 2006 in each county; and the number of students in sixth grade in the sample schools. This allowed us to create a set of weights to make our results more representative of Shaanxi province. With these weights, students in small schools and in small counties (small, according to the number of students) are given smaller weights; students in large schools and in large counties are given greater weight in the analysis.

are record-based (not from recall). In other words, the information on school achievement is not from recall, but is from each student's record book. The scores were copied by our enumerator with the assistance of the homeroom teacher.

In this paper, we focus on second term Chinese language scores because the scores for these classes are based on a single year-end test that is standardized. The exams are standardized in two dimensions. First, the questions are the same for all students within the schools in the same township. Second, the final exams were graded according to a single set of criteria by a township-wide panel of teachers (which is done to make the scores more objective). Although we also collected scores on math performance, for the sake of brevity we put findings using math scores in a series of appendix tables. In general, the findings in the paper (which uses Chinese language scores) would hold up if we had used math scores instead.

We also collected detailed information on the grade retention histories of each student. The students reported which grade they repeated. They also told us how many times that they repeated each grade. All this information is also available in their booklets and the enumerator (with the help of the homeroom teacher) was asked to verify the information as well. As it turns out, only about eight students repeated more than one grade. Because they were so special, these students were dropped from the analysis. Therefore, in our analysis we are looking exclusively at students that were retained for one year and comparing them to students that were not retained.

Even with standardized scores, one thing that we are worried about is that the effect of grade retention might be magnified or attenuated if a student was moved from a "fast or accelerated" class to a "slower" one after he or she was retained. If this were the case then the effect of grade retention on a student's scores might be confounded with the class placement decision. This is definitely not an issue in 67% of the schools that we surveyed since there was only one class per grade (meaning, there was no choice in terms of class placement). When we interviewed teachers and principals in the other schools (those with two or more classes per grade), we were told that it was a policy in rural elementary schools not to divide the classes into accelerated and/or slower ones. Most scholars familiar with rural education—especially education in poor areas in rural China—concurred with this observation. We also used statistical analysis to test whether the distributions of two or more classes in a single school were equal and our findings are consistent with observations and interviews in the field—there is no fast-tracking of students in rural elementary schools in our sample areas.[2]

Another issue on grade retention that we are concerned about is on what criteria were the grade retention decisions made. In particular, it is important to know if grade retention is a process that is mostly based on rules set by the school or if it is mostly a process that is in the hands of parents. Although during our survey and fieldwork this ended up being a difficult question to ask and get consistent answers, we believe that the survey results clearly support the conclusion that the grade retention decision is mostly in the hands of the school authorities, mostly is based on rules and only in a minority number of cases is subject to negotiations between parents and teachers/school administrators. Almost 100% of teachers and school administrators that were surveyed replied that grade retention was rule based and not subject to negotiations with parents. It is easy in the case of many schools (and/or school districts) to find written rules for grade retention posted in the school office and in school files. In addition, more than 60% of the parents of children that were retained told us the same thing. Moreover, although around 40% of the parents of students that were retained said that they were involved in the decision to retain their child, in fact, when looking at the scores of their children, in all but a small fraction of the cases their children's scores were so low such that they should have been retained on the basis of school rules. Hence, it may be that although parents may have believed they played a role in their child's retention decision, the final decision may have turned out the same whether the parent had visited the school or not. There are very few cases that a parent requested his/her child be retained when his/her scores was sufficiently high (that is, above the failing cutoff line).

In addition to school achievement and grade retention information, we also included information in the survey that could be used to create variables to control for other observed factors that might be expected to affect school achievement (for use as control variables). Two sets of variables were collected. In a set of questions about student characteristics, we collected information about each student's gender, age and asked them whether or not they were student cadres. Student cadres are students in classes that are assigned (mostly by their teachers) as class leaders and are given responsibilities, such as maintaining the cleanliness of the classroom and collecting homework. The survey form also included questions on the characteristics of the student's parents and family. The dataset includes variables on each parent's age and education attainment as well as the household's land holdings and the total number of other household members.

3. Grade retention in poor areas in rural China

Based on our data, one of the results that stands out above all others is the high rate of grade retention in our sample schools. Out of the 1653 students in the sample schools, 35% of the students in rural primary school repeated at least one grade before they entered grade 6 (Table 1, row 6). If such high rates of retention are common throughout China, it is clear that in the mid-2000s the prohibition against retaining a maximum of 5% of students is no longer binding. In fact, references to high retention rates are increasingly common in the literature (Wang and Wang, 1999). For example, China Central Television reported that the retention rate in some elementary schools in Gansu province was as high as 30% (CCTV, 2006). These high rates reported in areas outside of our sample area imply that our data may well be capturing what is a fairly common phenomenon. Internationally, however, such high rates are less common. In the US, for example, the estimated grade retention rates of the students' aged 14 and under range from 6.69% to 1.23% between the first grade and the fifth grade (Eide and Showalter, 2001). Interestingly, in our sample more boy students (39%, column 2) were retained than girl students (30%, column 4).

Although the overall retention rate is high for primary school, in general, the rates at which students are asked to repeat grades vary over the six years of schooling. Clearly, the rate is highest for first grade. Fully 11% of first graders repeat their first year of elementary school (Table 1). Such a finding, however, is not special. In the US, for example, retention rates are almost always substantially higher in the first grade than in subsequent grades (Eide and Showalter, 2001).

[2] In order to test this proposition, in schools with two fifth grade classes (eight out of 36 schools in total) and three fifth grade classes (four out of 36 schools) we used kernel distribution plots to graph the distributions of the scores and compared them with each other. A visual comparison of the distribution of the scores among the classes within the same school showed that, indeed, the distributions of the classes appeared similar. To confirm this statistically, we used a two-sample Kolmogorov–Smirnov test to determine if the distributions of the scores between any two classes in the same school were equal. In all but four schools, we cannot reject the hypothesis that any two classes in the same school have the same score distributions. In summary, it appears as if ex-retention class placement bias is not an issue impacting our analysis.

Table 1
Summary statistics of grade retention rate by gender and grade.

	Repeated grade	Boys		Girls		Total	
		(1)	(2)	(3)	(4)	(5)	(6)
		Number	Retention rate	Number	Retention rate	Number	Retention rate
(1)	Grade 1	910	0.13	743	0.09	1653	0.11
(2)	Grade 2	**910**	**0.10**	**743**	**0.07**	**1653**	**0.09**
(3)	Grade 3	**910**	**0.07**	**743**	**0.05**	**1653**	**0.06**
(4)	Grade 4	**910**	**0.05**	**743**	**0.05**	**1653**	**0.05**
(5)	Grade 5	910	0.04	743	0.02	1653	0.03
(6)	Total	910	0.39	743	0.30	1653	0.35

Table 2
Probit regression analysis of the determinants of grade retention in rural China.

Dependent variable: grade retention dummy, =1 if the student repeated and 0 otherwise[a]				
		(1)	(2)	(3)
(1)	Gender dummy, =1 if the student is male and 0 otherwise	0.182 (1.91)*	0.159 (2.12)**	0.012 (0.15)
(2)	Starting age	−0.411 (6.38)***	−0.107 (2.33)**	0.040 (0.76)
(3)	Average score (second term) in 2002		−0.030 (8.46)***	−0.031 (8.24)***
(4)	Sibling dummy, =1 if the student has no sibling in 2002 and 0 otherwise	−0.066 (0.61)	−0.093 (1.07)	−0.091 (0.92)
(5)	Age of the father, year	0.027 (2.32)**	0.033 (3.48)***	0.021 (2.11)**
(6)	Education level of the father, years of schooling	−0.037 (1.70)*	−0.035 (1.99)**	−0.026 (1.33)
(7)	Education level of the mother, years of schooling	0.011 (0.57)	−0.024 (1.53)	−0.041 (2.37)**
(8)	Household total land holding, mu	−0.007 (0.59)	−0.001 (0.13)	0.001 (0.10)
(9)	Dummy, =1 if the value of the house is larger than 5000 yuan	0.124 −0.066	−0.124 −0.093	−0.209 −0.091
(10)	School_dummy	Yes	Yes	Yes
(11)	Observations	1527	1588	1590

Z statistics in parentheses.
[a] In model (1), grade retention dummy equals to 1 if the student repeated in grade 1 and 0 otherwise; In model (2), grade retention dummy equals to 1 if the student ever repeated a grade between grades 1 and 5; and 0 otherwise; and in model (3), grade retention dummy equals to 1 if the student ever repeated a grade between grades 2 and 4; and 0 otherwise.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

The other pattern in the retention data is that after grade 1 the retention rates fall steadily (Table 1). Between grade 2 and grade 4, the retention rate falls from 9% to 6% to 5%.[3] By the fifth grade, only 3% of students were retained. Interestingly, of the nearly 600 students that repeated grades, only eight of them repeated more than one grade.

So who are these students that were retained for at least one grade? To answer this, we ran a Probit regression to examine the determinants of repeating a grade (Table 2).[4] In other words, on the left hand side we included a dummy variable that equals one if the student was retained (and zero otherwise); on the right hand side we included a series of student, school and parent characteristics. We repeated the regressions with three alternative dependent variables, depending on if the student was retained in the first grade or not (column 1); if the student was ever retained in the second to fourth grades (column 3); and if the student was ever retained during any year in which she/he was in elementary school (column 2).

According to our descriptive regression results, we find that certain types of students tend to repeat grades more often than others, although the results differ between the regressions that include first grade repeaters (columns 1 and 2) and those that do not include them (column 3). For example, we find that, ceteris paribus, young boy students are more likely to be retained than young girl students in the first grade (row 1, column 1). Looking at students' entire time at elementary school (including the first grade) boy students are also more likely to be retained for one of the grades (row 1, column 2). The gender effect, however, is not observed during grades 2–4. Also, the age at which a child starts school is negatively associated with the tendency to repeat grades—especially for the first and second regressions (row 2, columns 1 and 2). However, like the gender effect, this correlation also disappears during grades 2–4 (column 3). In addition, the scores that students earned during the beginning year of elementary school are associated closely with whether those students repeated any grades (either grades 1–5—row 3, column 2; or grades 2–4—column 3). Perhaps not surprisingly, students that have higher grades in the first grade tend to have a lower probability of repeating a grade during the subsequent years. Finally, in all of the regressions, several of the other control variables (e.g., age of father) are robustly correlated with grade retention—regardless of the nature of the dependent variable.

3.1. Grade retention and school performance in poor areas in rural China

Most importantly, especially in our analysis, when students were retained, there was a relative improvement to their school

[3] In this paper we will mainly focus on the students that repeat grade 2, grade 3 and grade 4. We do so, since it is only for these students that we can compare the changes of their scores from before and after the year that they were retained. Unfortunately, since we do not have a grade before grade 1 and do not observe a grade after grade 5 (since we are surveying sixth graders), we cannot use these observations as part of our treatment group (since we can observe the effect of retention on grade change). In our sample those that were retained in grade 2, grade 3 and grade 4 accounted for 57% of all the students who had ever been retained; this accounts for 20% of the entire sample (and a higher amount of the usable sample, since we drop those that were retained during grade 1 and grade 5 from the analysis (they are not part of either the treatment or control group).

[4] It should be noted that the purpose of running this regression is for purely descriptive reasons—to see what factors are correlated with the tendency for an individual to be retained. We are not at all trying to assign causation.

Table 3
Differences of average scores of Chinese language and math courses (second term) between students that repeated a grade and those that did not repeat a grade.

Repeated grade			(1) Grade 1	(2) Grade 2	(3) Grade 3	(4) Grade 4	(5) Grade 5
(1)	Grade 1	Not Repeated (control)	72.96	72.49	70.10	69.13	70.99
(2)		Repeated (treatment)	72.28	71.55	68.94	67.88	69.01
(3)		Difference	−0.68	−0.94	−1.17	−1.25	−1.99
(4)	Grade 2	Not repeated (control)	*73.57*	**72.64**	*70.64*	69.63	71.46
(5)		Repeated (treatment)	*67.35*	**66.54**	*65.12*	64.65	66.80
(6)		Difference	*−6.22*	**−6.11**	*−5.53*	−4.98	−4.66
(7)	Grade 3	Not repeated (control)	74.00	*73.53*	**70.97**	*69.71*	71.62
(8)		Repeated (treatment)	68.12	*68.40*	**66.56**	*68.19*	69.12
(9)		Difference	−5.89	*−5.13*	**−4.41**	*−1.52*	−2.50
(10)	Grade 4	Not repeated (control)	74.38	73.89	*71.25*	**69.98**	*71.76*
(11)		Repeated (treatment)	68.45	68.26	*66.78*	**65.80**	*69.58*
(12)		Difference	−5.93	−5.63	*−4.47*	**−4.18**	*−2.19*
(13)	Grade 5	Not repeated (control)	74.60	74.03	71.41	70.17	71.81
(14)		Repeated (treatment)	70.17	71.39	68.07	65.94	70.69
(15)		Difference	−4.44	−2.64	−3.34	−4.23	−1.13

performance. Based on descriptive statistics using our data, when we examine changes in the scores of students before and after they were retained in the second grade, the gap in the scores is narrowing between the students who were retained during their second grade years and those that were not retained. It is true that the scores of those that were retained were lower than those that were not retained—both before and after the year that students repeated. However, their scores, on average, were 6.2 points lower before they were retained and only 5.5 points lower after they were retained.

This same pattern holds for those students that were retained in the third and fourth grades (Table 3). For the students retained in the third grade, the gap in the scores (between the year before and the year after retention) dropped from 5.1 to 1.5 (row 9, columns 2 and 4). For the students retained in the fourth grade, the gap dropped from 4.5 to 2.2 (row 12, columns 2 and 4). The implication of these findings (should they hold up in the multivariate analysis—see below) is that grade retention appears to be helping students by improving their scores in a relative sense.

The narrowing gap is also fairly robust in several dimensions. For example, the narrowing of the gap is found to persist over time at the level of primary school education. In other words, the gap in the scores of grade 5 between those that were retained in grade 2 (grade 3) and those that were not retained was narrower than the gap in the scores in grade 1 (grade 2). In addition, the falling gap shows up when we look at Chinese language scores and math scores (see Tables A1 and A2). To show this it can be seen that for the students retained in the second grade, the gap in the scores between them and those not retained dropped from 6.2 to 6.0 in Chinese language and from 6.3 to 5.1 in math (see Tables A1 and A2, row 6, columns 1 and 3).

In short, then, the descriptive results show that grade retention may be helping students. Although those that were retained have scores lower than those that were not retained, the gap is narrowing over time. Such a finding would mean that something (for example, allowing students a chance to catch up or allowing them to mature age-wise) is helping contribute positively to the school performance of individuals. However, it is important to remember that our results to this point are descriptive. It is possible that when other factors are held constant, this positive result will disappear. We also do not know if the point estimate is positive or if it is statistically significant or not (that is, it could be statistically equal to zero). In fact, the education literature contains many papers that discuss the tendency of other factors to affect scores. For example, one paper finds that girl students outperform the boy students (ERIC Development Team, 2001) in reading and writing in some grades. Other papers have found that the starting age of a student also affects school performance (Fredriksson and Öckert, 2005). Because of these effects (and possible interactions between them and retention and scores), multivariable analysis is needed to more fully explore the impacts of grade retention on the school performance.

4. Methodology

The objective of this part of the study is to examine the effect of grade retention on educational performance (Chinese language scores). In order to evaluate the effects of grade retention, conceptually we are making grade retention the treatment. In other words, our sample students are divided into a treatment group (those that were retained by the school and had to repeat a grade) and a comparison group (those that never repeated a grade). To do this, we employ a differences-in-differences estimation approach (DID). Using the DID approach allows us to compare the outcomes before and after a student repeated a grade with students not affected by the treatment (those who were not retained). By comparing the before–after change of treated groups with the before–after change of comparison groups, any common trends, which will show up in the outcomes (Chinese language scores) of the comparison groups as well as the treated groups, will be differenced out (Smith, 2004).[5]

In addition to the standard DID estimator, we implement three other DID estimators: an "unrestricted" version that includes the lagged dependent variable (the Chinese language score from the student's first year in the primary school) as a right hand variable,

[5] Since we use differencing, it is possible that our results are affected by a common phenomenon in statistical analysis called "regression to the mean." This is a phenomenon that is potentially common in education because student test scores are in part due to ability and in part due to random error (or chance). Therefore, if we are looking at the changes in scores of only the students with the worst scores in class, it is possible that their scores will improve, not from the intervention (which in this case is grade retention) but because many of the unlucky students on the first test will be luckier on the second test and it will appear as if they got better due to the intervention when in fact there was no effect of the intervention and there was only a naturally occurring regression to the mean. Of course, it could be that regression to the mean is more evident for one of our groups—e.g., those students that were retained. Because of this, we also used matching, which is a non-experimental way to create treatment and control groups that overcome the potential problems of regression to the mean. The results in our paper when we used either DID or matching were almost the same. Therefore, we do not believe our results are merely picking up regression to the mean effects.

an "adjusted" version that includes other covariates in addition to the treatment variable (in our case they are a series of control variables from 2002 or the pre-retention period), and an unrestricted/adjusted model that combines the features of both the "unrestricted" and "adjusted" model. In summary, the models to be estimated are:

Model (1), restricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \varepsilon_i$
Model (2), restricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \beta' \mathbf{X}_i + \varepsilon_i$
Model (3), unrestricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \varepsilon_i$
Model (4), unrestricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta \mathbf{X}_i + \varepsilon_i$

where i is an index for the student, $\Delta Score_i$ is the change of the second term Chinese language score of student i between the Chinese language score after the student was retained and the Chinese language score before the student was retained,[6] $RETAIN_i$ is the treatment variable (which makes δ the parameter of interest) and the $Score\_beforeretain_i$ is the score of the student for the grade of the year before the student was retained. Finally, the term $\mathbf{X}_i$ is a vector of covariates that are included to capture the characteristics of students, parents and households. If not stated otherwise, $\mathbf{X}_i$ also includes a set of 11 town indicator or dummy variables.[7]

4.1. Alternative estimation approaches

As might be expected, the effectiveness of DID depends on the validity of the assumption of "parallel trends."[8] The reality of our question (understanding the effect of grade retention on the scores of students) may mean that even though we control for a large number of observable variables in 2002 in the adjusted and unrestricted versions of the DID estimates, there could be other unobservable factors that may compromise this assumption. Because of the potential existence of other differences between students retained and students not retained, we also use propensity score matching (PSM), which is an approach that does not require the parallel trend assumption. PSM allows the analyst to match the treated and the comparisons when observable characteristics of students, who were retained, and observable characteristics of students, who were not retained, are continuous (Rosenbaum and Rubin, 1985).

In addition, to eliminate the bias due to time-invariant unobservable differences between retained students and non-retained students, we extend the cross-sectional PSM approach to a longitudinal setting and implement a differences-in-differences matching (DIDM) strategy. With DIDM we can exploit the data on the retained students in the grade before they repeated to construct the required counterfactual, instead of just using the data in a grade after they repeated (as was used in the traditional PSM analysis—which was describe above). The advantage of DIDM is that the assumptions that justify DIDM estimation are weaker than the assumptions necessary for DID or the conventional PSM estimator.

Although the above matching methods can significantly improve the reliability of matching estimators, producing results that have been shown to be very close to those based on a randomized design, statisticians counsel that geographic mismatch between matched observations should be avoided (Smith and Todd, 2005; Abadie and Imbens, 2006). In our case when we use PSM, even if we have added a set of township dummies when estimating the propensity scores, students that are from different townships, but that have similar propensity scores, may still be matched as a pair of treatment and comparison observations. Abadie and Imbens (2006) propose a method to eliminate the bias caused by imprecise matching of covariates between treatment and comparison observations using nearest neighbor matching.[9]

In making specific choices about the methodology, our approach is to minimize potential bias whenever possible. To minimize geographic mismatch, we enforce exact matching by township.[10] To do this, each treatment observation is matched to three comparison observations with replacement, which is few enough to enable exact matching by township for nearly all observations, but enough to reduce the asymptotic efficiency loss significantly (Abadie and Imbens, 2006). When we use this method for matching, we report our results as *multi-dimensional matching* results to differentiate this approach to matching from the traditional or *basic matching* approach that we also use (which was described above).[11] This approach has been shown to prevent the estimates from relying too heavily on just a few comparison observations. In other words, because we are not sure what is the best approach, apriori, we use all of the approaches and hope that our results are the same—regardless of the exact approach adopted.

[6] In our analysis, when we examine the short-term effect of grade retention on the student's school performance, $\Delta Score_i$ is the change of the second term score of student i between the grade just after the student was retained and the grade right before the student was retained. For example, in the case of the effect of grade retention in grade 2, $\Delta Score_i$ is the final grade from the third grade minus the final grade from the first grade; while when we examine the long-term effect of grade retention on the student's school performance, $\Delta Score_i$ is the change of the second term score of student i between the fifth grade and first grade.

[7] Although it is true that different townships have different exams and different grading criteria, in most of our regressions, especially in Regression Model 4 (the model that we ultimately rely on for drawing the overall conclusions for the paper), we also include a set of township indicator (or dummy) variables (one for every township). This means that what we are actually measuring is the average effect of grade retention across different townships. In other words, we do not use any of the inter-township variability in test scores in deriving the estimated retention effects.

In order to show this more clearly, we can rewrite model (4) as:

Model (4a): $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta' X_i' + \text{D-town} + \varepsilon_i$,

where β' is the same as β, but does not include the estimated coefficients associated with the 11 town dummy variables; where X_i is the same as X_i except the set of 11 town dummies (D-town) has been removed; where D-town is a set of 12 dummy variables. In fact, D-town is actually a matrix of 11 dummy variables, where D-town1 = 1, if observations belong to town 1 and zero otherwise; where D-town2 = 1, if the observations belong to town 2 and zero otherwise; and so on for D-town3 to D-town11. Since we do not include D-town12, this is the base township (measured by the intercept α). When running the model this way, the interpretation of δ is the average effect across the 12 sample townships of RETAIN on the change of the Score (or $\Delta Score_i$).

An alternative model that would lead to the exact same result would be to run the model:

Model (4b): $\Delta Score_i = \alpha + \delta' RETAIN_i + \gamma' Score\_beforeretain_i + \beta' X_i' + u_i$

for each town separately (that is run it 12 times). In this case, the average of the 12 estimated parameters (δ') from model (4b) = δ from model (4a).

[8] See details in the appendix for the description of the methodology used in this paper.

[9] They also developed a formula to estimate standard errors for matching with a fixed number of nearest neighbors that are asymptotically consistent and which can accommodate unobserved heterogeneity in the treatment effect. In this paper, we use the nearest neighbor matching algorithm with bias adjustment developed by Abadie and Imbens (2006).

[10] This is accomplished by assigning an arbitrarily high weight to the exact matching variable in defining the matching criteria.

[11] Matching is based on a set of covariates which are time-invariant or were measured in 2002. The weighting matrix uses the Mahalanobis metric, which is the inverse of the sample variance/covariance matrix of the matching variables. We chose a set of 11 matching variables (see Table 4) for household level matching. Furthermore, we use the propensity scores as a diagnostic tool to restrict the sample used in each matching estimation to those with common support. We also visually examined the graphs of the propensity scores and trimmed the sample if there was a large imbalance between control observations and treatment observations with similar propensity scores.

Table 4
Difference-in-differences analysis for the effect of grade retention on school performance of Chinese language[a].

Dependent variable: the change in the second term scores of Chinese language between grade 1 and grade 5					
		(1) Restricted and unadjusted	(2) Restricted and adjusted	(3) Unrestricted and unadjusted	(4) Unrestricted and adjusted
(1)	Grade *RETENTION* dummy, =1 if the student ever repeated in grade 2, 3 or 4 and 0 otherwise	3.337 (3.15)***	2.354 (1.94)*	0.043 (0.05)	−1.698 (1.75)*
(2)	Score before retention			−0.510 (16.45)***	−0.715 (19.81)***
(3)	Gender dummy, =1 if the student is male and 0 otherwise		0.938 (1.18)		−1.852 (2.96)***
(4)	The student's age in 2002, year		0.465 (0.89)		−0.707 (1.72)*
(5)	Student cadre dummy, =1 if the student was a cadre in 2002 and 0 otherwise		−3.083 (3.55)***		0.767 (1.12)
(6)	Mentor dummy, =1 if the student had a mentor in 2002 and 0 otherwise		−1.402 (1.09)		−0.639 (0.70)
(7)	Sibling dummy, =1 if the student has no sibling in 2002 and 0 otherwise		−0.273 (0.28)		−0.007 (0.01)
(8)	Age of the father, year		−0.109 (1.22)		0.006 (0.08)
(9)	Education level of the father, years of schooling		0.007 (0.03)		0.122 (0.75)
(10)	Education level of the mother, years of schooling		0.103 (0.53)		0.204 (1.32)
(11)	Household total land holding in 2002, mu		−0.032 (0.30)		0.002 (0.03)
(12)	Number of household members in 2002, person		−0.090 (0.23)		0.087 (0.28)
(13)	House value dummy, =1 if the value of the house is larger than 5000 yuan in 2002 and 0 otherwise		−0.787 (0.93)		−0.922 (1.46)
(14)	Town dummy		Yes		Yes
(15)	Observations	1396	1346	1396	1346
(16)	R^2	0.01	0.08	0.30	0.45

Robust t statistics in parentheses.

[a] The sample here excludes the students who repeated in grades 1, 5 and 6 and the regression models used in this table are the following specifications respectively:
Model (1), restricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \varepsilon_i$.
Model (2), restricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \beta \mathbf{X}_i + \varepsilon_i$.
Model (3), unrestricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \varepsilon_i$.
Model (4), unrestricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta \mathbf{X}_i + \varepsilon_i$,
where i is an index for the student, $\Delta Score_i$ is the change of the second term score of Chinese language of student i between the first grade and the fifth grade, $RETAIN_i$ is the treatment variable (which makes δ the parameter of interest) and the $Score\_beforeretain_i$ is the second term score of Chinese language of student i for the first grade. Finally, the term $\mathbf{X}_i$ is a vector of covariates that are included to capture the characteristics of the student, his/her parents and household.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

5. Results of multivariate analysis

The results of our DID analysis using the restricted specifications (that is, models (1) and (2)) demonstrate that the findings of the multivariate analysis are consistent with the descriptive analysis.[12] For example, when we use the restricted and unadjusted specification of the empirical model (Table 4, column 1), the results show that, ceteris paribus, the Chinese language scores of the students that were retained in grade 2, grade 3 and grade 4 rose relatively more than those students who never repeated a grade during this period of elementary school (row 1). The coefficient on the variable of interest is statistically significant. This finding (*from the most simple model*) suggests that grade retention actually improves the Chinese language performance of students that were retained, the same finding as that of the descriptive statistics that were reported in Table 3. This result does not change much when we use the restricted and adjusted specification (which is the same specification as in column 1, but also controls for a number of observable covariates—column 2, row 1). If these results were to hold up throughout the rest of the paper, we might conclude that there is actually a benefit that is accruing to students from the recent relaxation of restrictions on the maximum number of students that can be retained in a single year.

When we use the unrestricted specification (either the unadjusted or adjusted version of the model—that is model (3) or (4)), however, the results change sharply (Table 4, columns 3 and 4). By controlling for the Chinese language performance of the students when they were in grade 1 (or the year before any of the students were retained—which is accomplished by including the variable, $Score\_grade1_i$), neither of the signs on the coefficient of the grade retention variable during grade 2, grade 3 and grade 4 is significant positive (row 1, columns 3 and 4). In fact, the coefficient of interest is negative and significant in the model that includes both $Score\_grade1_i$ and the other covariates (or the unrestricted and adjusted model, column 4, row 1). In general, this result demonstrates that the Chinese language scores of students that repeated a grade (either grade 2 or grade 3 or grade 4), in fact, dropped relative to the scores of those students that had never repeated a grade. Therefore, the most important finding in Table 4 is that—at least for the unrestricted model—we can reject the hypothesis that grade retention improves school performance in Chinese language.[13]

[12] In the paper for brevity, we report all of the results for the scores on Chinese language exams. We do this in Tables 4–7. While it is possible that we would get different answers if we ran the analysis separately on Math scores (instead of Chinese language scores), in fact, there is little difference between the results from the Math scores and the results from the Chinese language scores. To show this, we have added Appendix Tables A3–A6, and show the same results for Math scores in the Appendix as we do for the Chinese language scores.

[13] These results also show the importance of controlling for a student's ability (or, at least, the grades earned in grade 1). The t-ratios associated with the coefficient of the $Score\_grade1_i$ variable are very high.

Table 5
Difference-in-differences analysis for the short-term effect of grade retention on school performance of Chinese language[a].

Dependent variable: the change in the second term scores of Chinese language right before and after the student repeated							
		Grade 2		Grade 3		Grade 4	
		(1)	(2)	(3)	(4)	(5)	(6)
		Unrestricted and unadjusted	Unrestricted and adjusted	Unrestricted and unadjusted	Unrestricted and adjusted	Unrestricted and unadjusted	Unrestricted and adjusted
(1)	Grade *RETENTION* dummy =1 if the student repeated in a certain grade and 0 otherwise[e]	−2.519 (2.01)**	−2.569 (2.05)**	0.203 (0.15)	−1.304 (1.01)	2.053 (1.51)	2.083 (1.55)
(2)	Score before retention[b]	−0.503 (17.25)***	−0.645 (18.75)***	−0.486 (18.73)***	−0.628 (18.75)***	−0.361 (13.11)***	−0.566 (17.12)***
(3)	Gender dummy =1 if the student is male and 0 otherwise		−1.093 (1.86)*		−2.319 (3.99)***		−1.805 (3.00)***
(4)	The student's age in 2002, year		−1.254 (3.50)***		−0.660 (1.79)*		−0.889 (2.40)**
(5)	Student cadre dummy, =1 if the student was a cadre in 2002 and 0 otherwise		1.576 (2.43)**		0.788 (1.22)		0.985 (1.66)*
(6)	Mentor dummy, =1 if the student had a mentor in 2002 and 0 otherwise		−1.490 (1.68)*		−1.520 (1.75)*		0.260 (0.30)
(7)	Sibling dummy, =1 if the student has no sibling in 2002 and 0 otherwise		−0.480 (0.64)		0.911 (1.24)		0.136 (0.18)
(8)	Age of the father, year		0.123 (1.58)		0.028 (0.40)		−0.037 (0.52)
(9)	Education level of the father, years of schooling		0.111 (0.83)		−0.042 (0.32)		0.105 (0.74)
(10)	Education level of the mother, years of schooling		0.177 (1.47)		0.367 (2.74)***		0.163 (1.13)
(11)	Household total land holding in 2002, mu		−0.124 (1.84)*		0.005 (0.07)		0.056 (0.85)
(12)	Number of household members in 2002, person		−0.048 (0.17)		0.381 (1.36)		0.140 (0.47)
(13)	House value dummy, =1 if the value of the house is larger than 5000 yuan in 2002 and 0 otherwise		−1.251 (2.01)**		−0.103 (0.17)		−0.344 (0.57)
(14)	Town dummy		Yes		Yes		Yes
(15)	Observations	1396	1346	1395	1345	1396	1346
(16)	R^2	0.32	0.42	0.30	0.38	0.17	0.32

Robust *t* statistics in parentheses.

[a] The sample here excludes the students who repeated in grades 1, 5 and 6 and the regression models used in this table are the following specifications respectively:
Models (1) and (3), unrestricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \varepsilon_i$.
Models (2) and (4), unrestricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta \mathbf{X}_i + \varepsilon_i$,
where *i* is an index for the student, $\Delta Score_i$ is the change of the second term score of Chinese language of student *i* between the grade right after the student was retained and the grade right before the student was retained. That is, for grade 2 in columns (1) and (2), $\Delta Score_i$ is the change in the second term scores of Chinese language between grade 1 and grade 3; for grade 3 in columns (3) and (4), $\Delta Score_i$ is the change in the second term scores of Chinese language between grade 2 and grade 4; for grade 4 in columns (5) and (6), $\Delta Score_i$ is the change in the second term scores of Chinese language between grade 3 and grade 5. $RETAIN_i$ is the treatment variable (which makes δ the parameter of interest) and the $Score\_beforeretain_i$ is the second term score of Chinese language of student *i* for the grade of the year before the student was retained. Finally, the term $\mathbf{X}_i$ is a vector of covariates that are included to capture the characteristics of student *i*, his/her parents and household.

[b] The score before retention is the score of Chinese language in the year right before the student repeated, that is, the second term score of Chinese language in 2002 for models (1) and (2), the second term score of Chinese language in 2003 for models (3) and (4) and the second term score of Chinese language in 2004 for models (5) and (6).

[*] Significant at 10%.

[**] Significant at 10%.

[***] Significant at 10%.

Table 6
Difference-in-difference analysis for the long-term effect of grade retention on school performance of Chinese language[a].

Dependent variable: the change in the second term scores of Chinese language between grade 1 and grade 5					
		Grade 2		Grade 3	
		(1) Unrestricted and unadjusted	(2) Unrestricted and adjusted	(3) Unrestricted and unadjusted	(4) Unrestricted and adjusted
(1)	Grade *RETENTION* dummy, =1 if the student repeated in a certain grade and 0 otherwise	−1.154 (0.97)	−2.397 (1.87)*	−0.801 (0.55)	−2.476 (1.83)*
(2)	Score before retention	−0.514 (16.36)***	−0.712 (20.14)***	−0.448 (14.70)**	−0.643 (17.49)**
(3)	Gender dummy, =1 if the student is male and 0 otherwise		−1.792 (2.88)***		−1.853 (2.98)***
(4)	The student's age in 2002, year		−0.822 (2.13)**		−0.761 (2.03)**
(5)	Student cadre dummy, =1 if the student was a cadre in 2002 and 0 otherwise		0.826 (1.22)		0.481 (0.71)
(6)	Mentor dummy, =1 if the student had a mentor in 2002 and 0 otherwise		−0.637 (0.70)		−0.455 (0.51)
(7)	Sibling dummy, =1 if the student has no sibling in 2002 and 0 otherwise		−0.060 (0.08)		0.301 (0.39)
(8)	Age of the father, year		0.004 (0.05)		0.039 (0.52)
(9)	Education level of the father, years of schooling		0.124 (0.77)		0.108 (0.69)
(10)	Education level of the mother, years of schooling		0.210 (1.37)		0.240 (1.56)
(11)	Household total land holding in 2002, mu		0.001 (0.01)		0.011 (0.14)
(12)	Number of household members in 2002, person		0.080 (0.26)		0.086 (0.28)
(13)	House value dummy, =1 if the value of the house is larger than 5000 yuan in 2002 and 0 otherwise		−0.940 (1.48)		−0.778 (1.24)
(14)	Town dummy		Yes		Yes
(15)	Observations	1396	1346	1393	1343
(16)	R^2	0.30	0.45	0.24	0.37

Robust *t* statistics in parentheses.
[a] The sample here excludes the students who repeated in grades 1, 5 and 6 and the regression models used in this table are the following specifications respectively:
Models (1) and (3), unrestricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \varepsilon_i$.
Models (2) and (4), unrestricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta \mathbf{X}_i + \varepsilon_i$,
where i is an index for the student, $\Delta Score_i$ is the change of the second term score of Chinese language of student i between the first grade and the fifth grade, $\Delta Score_i$ is the change in the second term scores of Chinese language between grade 1 and grade 5, $RETAIN_i$ is the treatment variable (which makes δ the parameter of interest) and the $Score\_beforeretain_i$ is the second term score of Chinese language of student i for the first grade. Finally, the term $\mathbf{X}_i$ is a vector of covariates that are included to capture the characteristics of the student, his/her parents and household.
* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

Moreover, the higher adjusted R^2 statistics in models that include grade 1 scores show that the unrestricted versions of the model (columns 3 and 4) fit the data better. In other words, when analyzing the effect of grade retention on school performance it is important to control for a student's ability (or his/her beginning scores). Therefore, in the rest of the paper, we focus on the unrestricted models.

The same basic results hold when we look at the *short-term effects* of grade retention in grade 2 or grade 3 or grade 4 (Table 5).[14] Whether using the Unadjusted version of the model (columns 1, 3 and 5) or the adjusted version of the model (columns 2, 4 and 6), we can not find any significant positive effect of grade retention on school performance. This is true for those that repeat grade 2 (columns 1 and 2), grade 3 (columns 3 and 4), and grade 4 (columns 5 and 6). In other words, our results are consistent with those in the international literature that raise concerns that grade retention is not beneficial to the average student (Holmes, 1989; Fine, 1991).

The results remain almost the same when we examine the long-term effects of grade retention in grade 2 and grade 3. In this paper, the long-term effect of grade retention is defined as the change in the Chinese language score of a student between grade 1 and grade 5. This means that we are measuring a three-year effect in the case of those students that were retained in grade 2 and a two-year effect in the case of those students that were retained in grade 3. When doing so, the results remain consistent and show that there is no positive long-term effect of grade retention (Table 6). This is true if the student repeated grade 2 (columns 1 and 2) or grade 3 (columns 3 and 4). It also is true regardless of the version of the model that we run. In fact, for those students that repeated either in grade 2 or grade 3, their scores of Chinese language not only did not rise, they actually dropped (significantly) by more than two points (columns 2 and 4).

The results of the analysis of the coefficients of some of the control variables in Tables 4 and 5 are both of interest (in and of themselves) and help to provide confidence that our data are relatively high quality (since many of them produce signs that are reasonable). For example, in all of the regressions in which we control for the beginning scores of the students, the significant negative signs of the coefficients on the gender variables suggest that there is a tendency for boys in China's primary schools to score lower than girls. The significant negative sign on the coefficient of the starting-age-of-the-student variable shows that the scores of older students drop relatively more than those of younger students. This finding is reasonable since students that enter primary school at an older age may have an initial advantage because they may be relatively more mature (Fredriksson and Öckert, 2005). The initial advantage, however, gradually disappears as younger children catch up over the course of primary school. The positive signs (and level of statistical significance in some of the regressions) of the coefficients on the level-of-education-of-the mother variable suggest that the education of the mother is an important determinant of the Chinese

[14] In this paper we define short-term effects as the grade of student that was retained in the year immediately following his/her year of retention. In other words, the short-term effect on students that were retained in grade 2 is seen by examining how, ceteris paribus, scores change between grades 1 and 3. Likewise, the short-term effect on students that were retained in grade 3 (or grade 4) is seen by examining how scores change between grades 2 and 4 (or grades 3 and 5).

Table 7
Propensity score matching and multi-dimensional matching estimators and the effect of grade retention on school performance of Chinese language of primary students in rural China[a].

Treatment variable		Propensity score matching		Differences-in-differences matching	
		Average treatment effect for the treated (1)	*t*-value/*z*-value[b]	Average treatment effect for the treated (2)	*t*-value/*z*-value[b]
Panel A. Short-term effect of grade retention					
RETAINED in grade 2					
(1a)	Basic matching	−1.44	(0.73)	−2.12	(1.22)
(1b)	Multi-dimensional matching	−4.52	(2.63)**	−4.52	(2.63)**
RETAINED in grade 3					
(2a)	Basic matching	−1.38	(0.66)	−0.69	(0.28)
(2b)	Multi-dimensional matching	0.71	(0.41)	3.66	(1.58)
RETAINED in grade 4					
(3a)	Basic matching	0.58	(0.24)	1.46	(0.69)
(3b)	Multi-dimensional matching	1.64	(0.92)	2.08	(0.93)
Panel B. Long term effect of grade retention					
RETAINED in grade 2					
(4a)	Basic matching	−0.75	(0.45)	−0.55	(0.28)
(4b)	Multi-dimensional matching	−3.20	(1.96)**	−3.20	(1.97)**
RETAINED in grade 3					
(5a)	Basic matching	−2.53	(1.07)	−1.61	(0.65)
(5b)	Multi-dimensional matching	−1.04	(0.57)	−1.04	(0.57)

[a] The method of nearest neighbor matching is used to get the basic matching results of propensity score matching and multi-dimension matching; and the covariates, $\mathbf{X}_i$, used in generating the propensity score estimates are the same as those in Table 4.
[b] *t*-values/*z*-values are reported in parentheses. *t*-values are calculated for the basic propensity score matching with the coefficients and standard errors which are bootstrapped using 1000 replications, and *z*-values are reported for the multi-dimensional matching.
** Significant at 1% level.

language scores of children. Somewhat surprisingly, the signs on the coefficients of the "did the child have a mentor" variable were mostly negative (and statistically significant in some of the regressions). While one might believe, ex ante, that mentoring should help improve grades, it is possible that it is precisely the students that have low (or falling) Chinese language scores that are those that need (and receive) mentoring. The signs of the coefficients on the student cadre variable (which was equal to one if the student was at some point during his/her elementary years appointed to a position of class leadership by the teachers) are positive, suggesting perhaps that teachers tend to turn to better students in making their assignments for class leadership.

Interestingly, the coefficients on variables such as number of siblings, land holdings and the value of the house are insignificant and the signs are unstable across the regressions models in Tables 4 and 5. While we cannot pinpoint the reason, it is possible that these variables could be measured with error (which would tend to force the coefficient towards zero). It is also possible that there is multicollinearity that is affecting the precision of the estimates of the coefficients of the right hand side variables.[15] It is also, of course, possible that the there are few true linkages in our sample between these variables and Chinese language scores.

[15] It is possible that multicollinearity among the right hand variables is affecting the precision of the estimates of the coefficients. In order to allay these fears, we have used the collinearity diagnostic package in Stata (coldiag2) to examine the condition number. The condition number of our right hand side set of variables (the **X** matrix) is 63, which, according to Besley et al. (1980), means that there is not very serious collinearity among the right hand side explanatory variables in the model (the condition number is 63, which is less than 100, a cutoff point that economists often believe is the dividing line between serious and not serious multicollinearity). Even if there was multicollinearity, there is no evidence using variance decomposition analysis that the collinearity involves the variables that have shifted between the different models (e.g., the siblings variable or the cadre variable). In fact, the largest variance component (of a single variable) that is associated with the eigenvector that has the highest condition index is only 46%. Therefore, there is no evidence that there is any serious collinearity that is affecting our results.

5.1. Results from alternative methods

The results of cross-sectional PSM analysis—regardless of the method of matching—also reveal that grade retention has no significant positive effect on the school performance of students. When examining the effect of grade retention on school performance for all the students who were ever retained in any grade (that is, either grade 2 or grade 3 or grade 4) using Basic Matching methods, there are no cases in which the coefficient on the treatment variable (*RETAIN*) is significant (Table 7, column 1, rows 1a, 2a, 3a, 4a, and 5a). The results remain almost the same when using Multi-dimensional Matching (column 1, rows 1b, 2b, 3b, 4b, and 5b) except for the students that repeated grade 2. That is, in the cases when the students repeated grade 3 or grade 4, the coefficient on the treatment variable (*RETAIN*) is insignificant and the in the case when the students repeated grade 2, the coefficient on the treatment variable (*RETAIN*) is negative and significant. Therefore, from the PSM analysis, we can reject the hypothesis that grade retention improves school performance as well.

Finally, the findings continue to remain largely consistent when using differences-in-differences matching (DIDM—Table 7, column 2). Regardless if we use Basic Matching (rows 1a, 2a, 3a, 4a, and 5a) or multi-dimensional matching (rows 1b, 2b, 3b, 4b, and 5b), none of the coefficients of the treatment variables are positive and significant. In fact, when using multi-dimensional matching, in the case of those students that were retained in grade 2, the coefficients are negative and significant.

Hence, whether using DID, PSM or DIDM, there is no evidence that grade retention in our sample of students has improved school performance. This is true if we look at the effect in the short- or long-run. In fact, there is some evidence that when students repeated grade 2, retention appears to have a negative effect on school performance. While we have no basis on which to determine the exact mechanism that is causing the fall in scores, it is consistent with an explanation that often appears in the international literature that suggests that when students are retained the fall in their self-

esteem, in fact, offsets any positive effect of allowing the student another year to catch up (Kellam et al., 1975).

6. Summary and conclusions

In this paper we have tried to understand whether or not grade retention helps or hurts school performance of the students that were retained for a year of schooling during their elementary school years. The issues have gained prominence since in recent years retention rates—at least anecdotally—have begun to rise. Policy makers—who at one time restricted retention rates to not exceed a maximum level—should want to know how school performance of children is being affected when local educators raise the frequency of grade retention. According to the international literature, it is possible that grade retention can either benefit students (by giving them time to mature and catch up) or hurt them (by harming self-esteem and/or removing them from their original set of peers).

According to the results in this paper, we show—perhaps somewhat surprisingly—that there is no positive effect of grade retention on school performance of the students that were retained. Whether in the short term (the year immediately after a student was retained) or longer term (by grade 5), we can reject the hypothesis that grade retention improves the scores of the students that were retained. This result is true for students that were retained in grade 2, grade 3 and grade 4. In fact, in the analysis of some students that were retained (especially those that were retained in grade 2) grade retention was shown to have a statistically significant and negative effect on school performance.

Based on these results, it is possible to conclude that the conscious or unconscious decision to relax the rule to limit retention rates to a maximum level (which was originally made by education officials to limit the use of scarce fiscal resources that were allocated for public education) has actually had little benefit for—and may have had negative effects on—the school performance of the sample students. It is unclear why retentions rates have risen in recent years. If, as some have suggested, the rise in retention rates is due to some unintended incentive of funding arrangements that allows local elementary schools to increase revenues when student enrollments are higher—including the participation of students that have been retained—there needs to be investigation into ways to curb such actions.

There are also other, more far-reaching actions that these results may be advocating. It is also possible that grade retention would have a more positive effect on students if there were more complementary educational services available—such as counseling, tutoring sessions or, at the very least, an effort made by schools to make grade retention a more positive thing—and try in some way to reduce the stigma that could lead to falling self-esteem. We understand that our current results are not rich enough to provide evidence on which any of these further actions could be justified. However, the paper does produce results that should lead to calls for further research efforts that can be designed to better understand the effect of grade retention on the school performance of rural children.

Acknowledgements

The authors would like to thank Renfu Luo and Weiwei Zhao, who have spent uncountable days coordinating the survey and cleaning data. A special thanks to all the enumerators, school principals and students. We are also grateful for the useful comments from the anonymous referees. We acknowledge grants to support field research from The Ford Foundation (Beijing), Chinese Academy of Sciences (KSCX2-YW-N-039) and support for follow-up research from the National Natural Science Foundation of China (70803047 and 70903064), the Natural Science Foundation of Zhejiang Province (Y607420), the Social Sciences Foundation of Zhejiang Province (07CGLJ005YBQ) and the Research Fund for the Youth in Zhejiang Gongshang University..

Appendix A

Description of the methodology used in this paper

A.1. Differences-in-differences estimation

The objective of this study is to examine the effects of the grade retention on the student's educational performance (Chinese language scores). In order to evaluate the effects of grade retention, conceptually we are making grade retention the treatment. In other words, our sample students are divided into a treatment group (those that were retained by the school and had to repeat a grade) and a comparison group (those that never repeated a grade). More specifically, the treatment group includes all the students who ever repeated the second grade, third grade or fourth grade. The comparison group includes all the students who were never retained, but does not include any students who were retained in either the first or fifth grade (which are dropped from the sample). With this setup, we are interested in understanding the mean impact of "treatment on the treated," which is the average impact of grade retention among those treated (Smith and Todd, 2005):

$$\begin{aligned} TT &= E((Y_1 - Y_0)|X, D = 1) \\ &= E(Y_1|X, D = 1) - E(Y_0|X, D = 1) \end{aligned} \tag{A.1}$$

where we denote Y_1 as the outcome (the Chinese language scores of students—in our case) after the student was retained and Y_0 as the outcome if a student was not retained. In Eq. (A.1), our treatment is denoted by $D = 1$ which stands for the students who were retained for at least one grade and for whom Y_1 is observed and $D = 0$ stands for those who were not retained for whom Y_0 is observed. Because in reality we do not observe either the counterfactual mean, $E(Y_0|X, D = 1)$, or the mean outcome for the students had they not been retained in a grade after they were retained, we need to employ a differences-in-differences estimation approach (DID). Using the DID approach allows us to compare the outcomes before and after a student repeated a grade with students not affected by the treatment (those who were not retained).

In Eq. (A.1) let t and t' denote time periods after and before the change of grade retention. When doing so, the standard DID estimate is given by:

$$\begin{aligned} DD = &[E(Y_t|D = 1) - E(Y_{t'}|D = 1)] \\ &- [E(Y_t|D = 0) - E(Y_{t'}|D = 0)] \end{aligned} \tag{A.2}$$

The idea of using a DID estimator to estimate the effect of the treatment on the treated is that it allows us to correct for the differences before and after the treatment (that is for the scores before and after a student was retained) by subtracting the simple difference for the comparison group (not retained students). By comparing the before – after change of treated groups with the before – after change of comparison groups, any common trends, which will show up in the outcomes (Chinese language scores) of the comparison groups as well as the treated groups, will be differenced out (Smith, 2004).

A.2. Alternative estimation strategies

It is important to remember that the statistical identification of the causal effects using DID relies on the assumption that absent the policy change (or grade retention in our case), the average change in the Chinese language scores ($Y_t - Y_{t'}$) would have been the same for the treated and the comparison groups. Formally, this is called the "parallel trend" assumption, which can be expressed as:

$$E(Y_{0,t}|D=1) - E(Y_{0,t'}|D=1) = E(Y_{0,t}|D=0) - E(Y_{0,t'}|D=0) \quad (A.3)$$

As might be expected, the effectiveness of DID depends on the validity of this assumption. The reality of our question (understanding the effect of grade retention on the scores of students) may mean that even though we control for a large number of observable variables in 2002 in the adjusted and unrestricted versions of the DID estimates, there could be other unobservable factors that may compromise the parallel trend assumption. Because of the potential existence of other differences between students retained and students not retained, we also use propensity score matching (PSM), which is an approach that does not require the parallel trend assumption. PSM allows the analyst to match the treated and the comparisons when observable characteristics of students, who were retained, and observable characteristics of students, who were not retained, are continuous (Rosenbaum and Rubin, 1985). With the right data, it is possible to estimate the propensity scores of all students and compare the match scores (or outcomes) of students who were retained and those who were not retained that have similar propensity scores.[16] In this way, then, we can obtain the mean impact of the treatment on the treated (Dehejia and Wahba, 2002; Smith and Todd, 2005):

$$E(Y_1 - Y_0|D=1) = E(Y_1|D=1) - E_{Z|D=1}\{E(Y_0|\,p(Z), D=0)\} \quad (A.4)$$

where $p(Z) \equiv Pr(D=1|Z)$ is the propensity score. Matching is based on the assumption that outcomes (Y_0, which is a Chinese language score of the student—in our case) are independent of participation (grade retention) conditional on a set of observable characteristics (Rosenbaum and Rubin, 1983). Because of this assumption, we do not need to worry about unobservable heterogeneity. By matching students who were retained and students who were not retained with similar values of, any differences in $E(Y_0)$ between the two groups are assumed to be differenced $Pr(D=1|Z)$ out when calculating the above equation. The assumption of matching is that $E(Y_0|Z, D=1) = E(Y_0|Z, D=0)$. The observable covariates Z should include the characteristics that determine grade retention. In our analyses,includes a number of Z variables including student, parent and household characteristics. We also include township fixed effects to control for unobservable factors at the township level that may affect grade retention.

To implement PSM successfully, however, the nature of the students who were retained and the nature of the students who were not retained must meet certain criteria. Importantly, the common support of propensity scores for participating and non-participating students should be fairly wide. Intuitively, wide common support means that there must be a fairly large overlap in the propensity scores between the treated and comparison groups. In our sample, the common support is fairly wide.[17] This means that we are able to estimate the *average* treatment effect for the treated of a large portion of the sample.[18]

To eliminate the bias due to time-invariant unobservable differences between retained students and non-retained students, we extend the cross-sectional PSM approach to a longitudinal setting and implement a differences-in-differences matching (DIDM) strategy. With DIDM we can exploit the data on the retained students in the grade before they repeated to construct the required counterfactual, instead of just using the data in a grade after they repeated (as was used in the traditional PSM analysis—which was describe above). The advantage of DIDM is that the assumptions that justify DIDM estimation are weaker than the assumptions necessary for DID or the conventional PSM estimator. DIDM only requires that in the absence of treatment, the average outcomes for treated and comparisons would have followed parallel paths:

$$E(Y_{0,t}|P(Z), D=1) - E(Y_{0,t'}|P(Z), D=1) = E(Y_{0,t}|P(Z), D=0) - E(Y_{0,t'}|P(Z), D=0) \quad (A.7)$$

Assumptions embedded in Eq. (A.7) are weaker than the assumptions necessary for DID. Intuitively, DIDM removes time invariant unobservable differences between retained students and non-retained students conditional on $P(Z)$, a clear advantage over cross-sectional PSM.[19]

Although the above matching methods can significantly improve the reliability of matching estimators, producing results that have been shown to be very close to those based on a randomized design counsel that geographic mismatch between matched observations should be avoided (Smith and Todd, 2005; Abadie and Imbens, 2006). In our case when we use PSM, even if we have added a set of township dummies when estimating the

[16] We need to note, however, that a recent study found that the propensity score matching method is sensitive to the covariates used to estimate the scores and that combination of matching with DD was superior (Smith, 2004). We account for this comment below.

[17] The results are available upon request.

[18] Once we determine that PSM is feasible, we next need to choose the method of matching. In our analysis, we choose to use the nearest neighbor matching method with replacement. Following Smith and Todd (2005), we match on the log odds-ratio and standard errors are bootstrapped using 1000 replications. We also use a balancing test that follows Dehejia and Wahba (1999, 2002) that is satisfied for all covariates. The results of the balancing tests are available upon request.

While PSM is often used in program evaluations, it relies on a key underlying assumption: outcomes are independent of grade retention conditional on a set of observable characteristics. Formally, this assumption can be written as:

$$E(Y_0|P(Z), D=1) = E(Y_0|P(Z), D=0) \quad (A.6)$$

In other words, there would be no need to worry about unobservable heterogeneity. However, even though we control for unobservable differences at the township level using fixed effects when estimating the propensity score, there may still be systematic differences between the outcomes of retained students and not-retained students. The systematic differences could arise, for example, because the student's decision to repeat his grade is based on some unmeasured household or personal characteristics. Such differences could violate the identification conditions required for matching (Smith and Todd, 2005).

[19] Using outcomes from experimental data as a benchmark, Smith and Todd (2004) found that DDM performed better than DD or PSM methods. In performing DDM we match by using the log odds-ratios and the same nearest neighbor matching methods with replacement that were used in our PSM approach (which were described above). In addition, we also compute the "adjusted" version where the control units are weighted by the number of times that they are matched to a treated unit. The standard errors also are bootstrapped using 1000 replications.

Although the above matching methods can significantly improve the reliability of matching estimators, producing results that have been shown to be very close to those based on a randomized design (Smith and Todd, 2005; Abadie and Imbens, 2006), Smith and Todd (2005) counsel that geographic mismatch between matched observations should be avoided. In our case, when we use PSM, even if we have added a set of township dummies when estimating the propensity scores, students that are from different townships but have similar propensity scores may still be matched as a pair of treatment and control observations. Abadie and Imbens (2006) propose a method to eliminate bias caused by imprecise matching of covariates between treatment and control observations using nearest neighbor matching.

propensity scores, students that are from different townships, but that have similar propensity scores, may still be matched as a pair of treatment and control observations. Abadie and Imbens (2006) propose a method to eliminate the bias caused by imprecise matching of covariates between treatment and comparison observations using nearest neighbor matching.[20]

In making specific choices about the methodology, our approach is to minimize potential bias whenever possible. To minimize geographic mismatch, we enforce exact matching by township.[21] To do this, each treatment observation is matched to three comparison observations with replacement, which is few enough to enable exact matching by township for nearly all observations, but enough to reduce the asymptotic efficiency loss significantly (Abadie and Imbens, 2006). This approach has been shown to prevent the estimates from relying too heavily on just a few control observations. In other words, because we are not sure what is the best approach, apriori, we use all of the approaches and hope that our results are the same—regardless of the exact approach adopted.

Table A1 Differences of the score of Chinese language (second term) between students that repeated a grade and those that did not repeat a grade.

	Repeated grade		(1) Grade 1	(2) Grade 2	(3) Grade 3	(4) Grade 4	(5) Grade 5
(1)	Grade 1	Not repeated (control)	72.53	72.11	70.12	69.26	69.36
(2)		Repeated (treatment)	71.04	70.80	69.57	67.57	66.99
(3)		Difference	−1.49	−1.31	−0.55	−1.69	−2.37
(4)	Grade 2	Not repeated (control)	*73.12*	**72.70**	*70.68*	69.74	69.78
(5)		Repeated (treatment)	*66.97*	**66.76**	*64.71*	64.79	65.55
(6)		Difference	*−6.15*	**−5.94**	*−5.97*	−4.95	−4.23
(7)	Grade 3	Not repeated (control)	73.57	*73.04*	**70.88**	*69.87*	70.00
(8)		Repeated (treatment)	67.83	*68.66*	**68.05**	*67.74*	66.88
(9)		Difference	−5.74	*−4.38*	**−2.83**	*−2.13*	−3.12
(10)	Grade 4	Not repeated (control)	74.02	73.40	*71.19*	**70.12**	*70.12*
(11)		Repeated (treatment)	66.99	67.80	*66.24*	**66.09**	*68.38*
(12)		Difference	−7.03	−5.60	*−4.95*	**−4.03**	*−1.74*
(13)	Grade 5	Not repeated (control)	74.27	73.56	71.32	70.33	70.22
(14)		Repeated (treatment)	68.90	70.29	68.79	65.79	68.03
(15)		Difference	−5.37	−3.27	−2.53	−4.54	−2.19

Table A2 Difference of the score of math (second term) between students that repeated a grade and those that did not repeat a grade.

	Repeated grade		(1) Grade 1	(2) Grade 2	(3) Grade 3	(4) Grade 4	(5) Grade 5
(1)	Grade 1	Not repeated (control)	73.39	72.86	70.08	69.00	72.62
(2)		Repeated (treatment)	73.51	72.29	68.30	68.19	71.02
(3)		Difference	0.12	−0.57	−1.78	−0.81	−1.60
(4)	Grade 2	Not repeated (control)	*74.01*	**72.58**	*70.60*	69.51	73.14
(5)		Repeated (treatment)	*67.72*	**66.31**	*65.52*	64.50	68.05
(6)		Difference	*−6.29*	**−6.27**	*−5.08*	−5.01	−5.09
(7)	Grade 3	Not repeated (control)	74.43	*74.01*	**71.05**	*69.54*	73.23
(8)		Repeated (treatment)	68.40	*68.14*	**65.07**	*68.64*	71.36
(9)		Difference	−6.03	*−5.87*	**−5.98**	*−0.90*	−1.87
(10)	Grade 4	Not repeated (control)	74.74	74.38	*71.30*	**69.83**	*73.40*
(11)		Repeated (treatment)	69.91	68.72	*67.32*	**65.50**	*70.77*
(12)		Difference	−4.83	−5.66	*−3.98*	**−4.33**	*−2.63*
(13)	Grade 5	Not repeated (control)	74.93	74.49	71.50	70.01	73.40
(14)		Repeated (treatment)	71.43	72.48	67.34	66.09	73.34
(15)		Difference	−3.50	−2.01	−4.16	−3.92	−0.06

Table A3 DinD analysis for the effect of grade retention on school performance of math[a].

Dependent variable: the change in the second term scores of math between grade 1 and grade 5					
		(1) Restricted and unadjusted	(2) Restricted and adjusted	(3) Unrestricted and unadjusted	(4) Unrestricted and adjusted
(1)	Grade *RETENTION* dummy, =1 if the student ever repeated in grade 2, 3 or 4 and 0 otherwise	1.579 (1.39)	0.519 (0.43)	−1.142 (1.18)	−3.314 (3.44)***
(2)	Score before retention			−0.473 (14.42)***	−0.688 (18.21)***
(3)	Gender dummy, =1 if the student is male and 0 otherwise		1.147 (1.42)		0.501 (0.78)
(4)	The student's age in 2002, year		−0.603 (1.15)		−1.441 (3.30)***
(5)	Student cadre dummy, =1 if the student was a cadre in 2002 and 0 otherwise		−1.442 (1.65)*		2.194 (3.02)***
(6)	Mentor dummy, =1 if the student had a mentor in 2002 and 0 otherwise		−2.578 (1.96)*		−2.045 (2.09)**
(7)	Sibling dummy, =1 if the student has no sibling in 2002 and 0 otherwise		0.858 (0.89)		0.946 (1.21)

[20] They also developed a formula to estimate standard errors for matching with a fixed number of nearest neighbors that are asymptotically consistent and which can accommodate unobserved heterogeneity in the treatment effect. In this paper, we use the nearest neighbor matching algorithm with bias adjustment developed by Abadie and Imbens (2006).

[21] This is accomplished by assigning an arbitrarily high weight to the exact matching variable in defining the matching criteria.

(*Continued*)

Dependent variable: the change in the second term scores of math between grade 1 and grade 5					
		(1) Restricted and unadjusted	(2) Restricted and adjusted	(3) Unrestricted and unadjusted	(4) Unrestricted and adjusted
(8)	Age of the father, year		−0.018 (0.16)		0.001 (0.01)
(9)	Education level of the father, years of schooling		−0.222 (1.02)		−0.059 (0.38)
(10)	Education level of the mother, years of schooling		0.153 (0.88)		0.233 (1.70)*
(11)	Household total land holding in 2002, mu		0.008 (0.08)		0.036 (0.46)
(12)	Number of household members in 2002, person		0.339 (0.88)		0.493 (1.56)
(13)	House value dummy, =1 if the value of the house is larger than 5000 yuan in 2002 and 0 otherwise		0.192 (0.21)		−0.071 (0.10)
(14)	Town dummy		Yes		Yes
(15)	Observations	1398	1348	1398	1348
(16)	R^2	0.00	0.14	0.20	0.42

Robust *t* statistics in parentheses.

[a] The sample here excludes the students who repeated in grades 1, 5 and 6 and the regression models used in this table are the following specifications respectively:
Model (1), restricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \varepsilon_i$.
Model (2), restricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \beta \mathbf{X}_i + \varepsilon_i$.
Model (3), unrestricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \varepsilon_i$.
Model (4), unrestricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta \mathbf{X}_i + \varepsilon_i$,
where i is an index for the student, $\Delta Score_i$ is the change of the second term score of math of student i between the first grade and the fifth grade, $RETAIN_i$ is the treatment variable (which makes δ the parameter of interest) and the $Score\_beforeretain_i$ is the second term score of math of the student for the first grade. Finally, the term $\mathbf{X}_i$ is a vector of covariates that are included to capture the characteristics of student i, his/her parents and household.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

Table A4 DinD analysis for the short-term effect of grade retention on school performance of math[a].

Dependent variable: the change in the second term scores of math right before and after the student repeated							
		Grade 2		Grade 3		Grade 4	
		(1) Unrestricted and unadjusted	(2) Unrestricted and adjusted	(3) Unrestricted and unadjusted	(4) Unrestricted and adjusted	(5) Unrestricted and unadjusted	(6) Unrestricted and adjusted
(1)	Grade *RETENTION* dummy =1 if the student repeated in a certain grade and 0 otherwise[e]	−2.793 (2.20)**	−1.851 (1.42)	0.322 (0.22)	0.945 (0.73)	−0.161 (0.10)	−0.776 (0.53)
(2)	Score before retention[b]	−0.481 (16.13)***	−0.601 (13.51)***	−0.603 (22.28)***	−0.650 (20.36)***	−0.388 (9.18)***	−0.595 (12.35)***
(3)	Gender dummy =1 if the student is male and 0 otherwise		0.071 (0.11)		−0.504 (0.84)		0.376 (0.59)
(4)	The student's age in 2002, year		−1.721 (4.55)***		−1.332 (3.37)***		−1.685 (3.90)***
(5)	Student cadre dummy, =1 if the student was a cadre in 2002 and 0 otherwise		2.737 (3.55)***		1.743 (2.89)***		2.178 (2.91)***
(6)	Mentor dummy, =1 if the student had a mentor in 2002 and 0 otherwise		0.046 (0.04)		−0.727 (0.77)		−1.813 (1.84)*
(7)	Sibling dummy, =1 if the student has no sibling in 2002 and 0 otherwise		−0.090 (0.11)		−0.503 (0.68)		0.896 (1.14)
(8)	Age of the father, year		0.057 (0.75)		0.100 (1.29)		−0.028 (0.30)
(9)	Education level of the father, years of schooling		0.322 (1.96)*		−0.071 (0.51)		−0.138 (0.86)
(10)	Education level of the mother, years of schooling		0.077 (0.57)		0.249 (1.76)*		0.230 (1.59)
(11)	Household total land holding in 2002, mu		−0.101 (1.39)		−0.020 (0.25)		0.079 (1.02)
(12)	Number of household members in 2002, person		0.229 (0.76)		0.169 (0.57)		0.410 (1.36)
(13)	House value dummy, =1 if the value of the house is larger than 5000 yuan in 2002 and 0 otherwise		0.364 (0.53)		0.351 (0.53)		−0.088 (0.13)
(14)	Town dummy		Yes		Yes		Yes
(15)	Observations	1398	1348	1391	1341	1397	1347
(16)	R^2	0.26	0.40	0.35	0.48	0.15	0.36

Robust *t* statistics in parentheses.

[a] The sample here excludes the students who repeated in grades 1, 5 and 6 and the regression models used in this table are the following specifications respectively:
Models (1) and (3), unrestricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \varepsilon_i$.
Models (2) and (4), unrestricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta \mathbf{X}_i + \varepsilon_i$,
where i is an index for the student, $\Delta Score_i$ is the change of the second term score of math of student i between the grade right after the student was retained and the grade right before the student was retained. That is, for grade 2 in columns (1) and (2), $\Delta Score_i$ is the change in the second term scores of math between grade 1 and grade 3; for grade 3 in columns (3) and (4), $\Delta Score_i$ is the change in the second term scores of math between grade 2 and grade 4; for grade 4 in columns (5) and (6), $\Delta Score_i$ is the change in the second term scores of math between grade 3 and grade 5. $RETAIN_i$ is the treatment variable (which makes δ the parameter of interest) and the $Score\_beforeretain_i$ is the second term score math of student i for the grade of the year before the student was retained. Finally, the term $\mathbf{X}_i$ is a vector of covariates that are included to capture the characteristics of student i, his/her parents and the household.

[b] The score before retention is the second term score of math in the year right before the student repeated, that is, the second term score of math in 2002 for models (1) and (2), the second term score of math in 2003 for models (3) and (4)and the second term score of math in 2004 for the in the models (5) and (6).

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

Table A5 DinD analysis for the long-term effect of grade retention on school performance of math[a].

Dependent variable: the change in the second term scores of math between grade 1 and grade 5					
		Grade 2		Grade 3	
		(1)	(2)	(3)	(4)
		Unrestricted and unadjusted	Unrestricted and adjusted	Unrestricted and unadjusted	Unrestricted and adjusted
(1)	Grade *RETENTION* dummy, =1 if the student repeated in a certain grade and 0 otherwise	−3.588 (2.42)**	−4.503 (3.19)***	0.940 (0.58)	−2.198 (1.51)
(2)	Score before retention	−0.476 (14.45)***	−0.681 (18.54)***	−0.455 (10.63)***	−0.670 (15.38)***
(3)	Gender dummy, =1 if the student is male and 0 otherwise		0.592 (0.92)		0.181 (0.28)
(4)	The student's age in 2002, year		−1.687 (4.20)***		−1.746 (4.00)***
(5)	Student cadre dummy, =1 if the student was a cadre in 2002 and 0 otherwise		2.317 (3.22)***		2.380 (3.19)***
(6)	Mentor dummy, =1 if the student had a mentor in 2002 and 0 otherwise		−2.041 (2.09)**		−2.097 (2.11)**
(7)	Sibling dummy, =1 if the student has no sibling in 2002 and 0 otherwise		0.845 (1.08)		0.623 (0.80)
(8)	Age of the father, year		−0.004 (0.04)		−0.012 (0.13)
(9)	Education level of the father, years of schooling		−0.055 (0.36)		−0.057 (0.34)
(10)	Education level of the mother, years of schooling		0.246 (1.82)*		0.280 (1.90)*
(11)	Household total land holding in 2002, mu		0.033 (0.43)		0.055 (0.68)
(12)	Number of household members in 2002, person		0.478 (1.51)		0.482 (1.54)
(13)	House value dummy, =1 if the value of the house is larger than 5000 yuan in 2002 and 0 otherwise		−0.091 (0.13)		0.086 (0.12)
(14)	Town dummy		Yes		Yes
(15)	Observations	1398	1348	1393	1343
(16)	R^2	0.21	0.42	0.19	0.41

Robust *t* statistics in parentheses.

[a] The sample here excludes the students who repeated in grades 1, 5 and 6 and the regression models used in this table are the following specifications respectively:
Models (1) and (3), unrestricted and unadjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \varepsilon_i$.
Models (2) and (4), unrestricted and adjusted: $\Delta Score_i = \alpha + \delta RETAIN_i + \gamma Score\_beforeretain_i + \beta \mathbf{X}_i + \varepsilon_i$,
where i is an index for the student, $\Delta Score_i$ is the change of the second term score of math of student i between the first grade and the fifth grade, $\Delta Score_i$ is the change in the second term scores of math between grade 1 and grade 5, $RETAIN_i$ is the treatment variable (which makes δ the parameter of interest) and the $Score\_beforeretain_i$ is the second term score of math of student i for the first grade. Finally, the term $\mathbf{X}_i$ is a vector of covariates that are included to capture the characteristics of student i, his/her parents and household.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

Table A6 Propensity sore matching and multi-dimensional matching estimators and the effect of grade retention on school performance of math of primary students in rural China[a].

Treatment variable		Propensity score matching		Differences-in-differences matching	
		Average treatment effect for the treated	*t*-value/*z*-value[b]	Average treatment effect for the treated	*t*-value/*z*-value[b]
		(1)		(2)	
Panel A. Short-term effect of grade retention					
RETAINED in grade 2					
(1a)	Basic matching	−0.20	(0.09)	0.15	(0.07)
(1b)	Multi-dimensional matching	−4.44	(3.01)***	−4.44	(3.01)***
RETAINED in grade 3					
(2a)	Basic matching	3.16	(1.47)	3.85	(1.31)
(2b)	Multi-dimensional matching	0.85	(0.43)	1.30	(0.52)
RETAINED in grade 4					
(3a)	Basic matching	0.06	(0.02)	0.65	(0.23)
(3b)	Multi-dimensional matching	−3.71	(1.73)*	−5.24	(2.19)***
Panel B. Long term effect of grade retention					
RETAINED in grade 2					
(4a)	Basic matching	−1.80	(0.82)	−3.22	(1.67)
(4b)	Multi-dimensional matching	−3.96	(1.99)***	−3.96	(1.99)***
RETAINED in grade 3					
(5a)	Basic matching	−2.74	(1.13)	0.33	(0.14)
(5b)	Multi-dimensional matching	−2.44	(1.17)	−2.44	(1.17)

[a] The method of nearest neighbor matching is used to get the basic matching results of propensity score matching and multi-dimension matching; and the covariates, $\mathbf{X}_i$, used in generating the propensity score estimates are the same as those in Table 4.

[b] *t*-values/*z*-values are reported in parentheses. *t*-values are calculated for the basic propensity score matching with the coefficients and standard errors which are bootstrapped using 1000 replications, and *z*-values are reported for the multi-dimensional matching.

* Significant at 10%.
** Significant at 5%.
*** Significant at 1%.

References

Abadie, A., Imbens, G., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74, 235–267.

Alexander, K.L., Entwisle, D.R., Dauber, S.L., 1994. On The Success of Failure. University of Cambridge Press, New York.

Besley, D.A., Welsch, R.E., Kuh, E., 1980. Regression Diagnostics, Identifying Influential Data and Sources of Collinearity. John Wiley & Sons Inc., New York, NY, pp. 85–116.

China Central TV (CCTV), 2006. The Amazing High retention Rate. 9.16: http://bbs.cctv.com/cache/forumbook.jsp?id=8914299&pg=1&agMode=1.

Dehejia, R.H., Wahba, S., 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. Journal of American Statistical Association 94, 1053–1062.

Dehejia, R.H., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. The Review of Economics and Statistics 84, 151–161.

Eide, E.R., Showalter, M.H., 2001. The effect of grade retention on educational and labor market outcomes. Economics of Education Review 20, 563–576.

ERIC development team, 2001. Gender Differences in Educational Achievement within Racial and Ethnic Groups. ED455341, http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/17/34/2d.pdf.

Fredriksson, P., Öckert B., 2005. Is Early Learning Really More Productive? The Effect of School Starting Age on School and Labor Market Performance. Working paper, IZA DP No. 1659.

Fine, M., 1991. Framing Dropouts: Notes on the Politics of an Urban Public High School. SUNY Press, Albany, NY.

Grissom, J.B., Shepard, L.A., 1989. Repeating and dropping out of school. In: Shepard, L.A., Smith, M.L. (Eds.), Flunking Grades: Research and Policies on Retention. Falmer, London, pp. 34–63.

Guangming Daily, 2000. Gansu will gradually abolish grade retention in the elementary schools. 3.28.

Holmes, C.T., 1989. Grade level retention effects: a meta-analysis of research studies. In: Shepard, L.A., Smith, M.L. (Eds.), Flunking Grades: Research and Policies on Retention. Falmer, London, pp. 16–33.

Huang, Zh, 1998. Retention shouldn't be abolished. Management of Elementary and Middle Schools 3.

Kellam, S.G., Branch, J.D., Agrawal, K.C., Ensminger, M.E., 1975. Mental Health and Going to School: The Woodlawn Program of Assessment, Early Intervention, and Evaluation. University of Chicago Press, Chicago, IL.

Kerzner, R.L., 1982. The Effect of Retention on Achievement. Kean College of New Jersey. Family, Home and Social Sciences at Brigham Young University, Union, NJ.

Li, H., 2004. Retention, a topic that should be readdressed. Jiaoyuzhongheng 3.

Liddell, C., Rae, G., 2001. Predicting early grade retention: a longitudinal investigation of primary school progress in a sample of rural South African children. British Journal Educational Psychology 71, 413–428.

Ministry of Education, 2006. The law of compulsory education in China. http://www.moe.edu.cn/edoas/website18/info20369.htm.

Rosenbaum, P., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rosenbaum, P., Rubin, D.B., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. American Statistician 39, 33–38.

Royce, J.M., Darlington, R.B., Murray, H.W., 1983. Pooled analysis: findings across studies. In: Consortium for Longitudinal Studies. As The twig is Bent: Lasting Effects of Preschool Programs, Erlbaum, Hillsdale, NJ, pp. 411–459.

Smith, J., 2004. Evaluating the Local Economic Development Policies: Theory and Practice. Unpublished, College Park, Maryland.

Smith, J., Todd, P., 2005. Does matching overcome Lalonde's critique of nonexperimental estimators? Journal of Econometrics 125, 305–353.

Wang, C., Zhou, F., 2005. Reforming fiscal policy after tax for fee reform in China. Working paper. World Bank, Washington, DC.

Wang, L., Wang J., 1999. Current situation of basic education in Northwest China, http://www.nwnu.edu.cn/yjzx/kcxzh.htm.

Wen, X., 2002. Promotion or retention, which is more efficient. Information of Education 5, 17–18.

Yang, N., 1991. Analysis on the dropping out and grade retention in the elementary and middle schools in China. People's Education 3.

Yu, Z., 1999. There should be a retention system in elementary schools. Management of Elementary and Middle Schools 9.