Grade Retention and Student Achievement – What is the impact over time?

Nadim Schumann

Ph.D. Student at University of Zurich

Abstract

Repetition rates in francophone sub-Saharan Africa are exceedingly high when compared to other world regions. These high rates pose a heavy burden on education systems struggling to achieve higher enrolment and better quality in primary education. Against the backdrop of the Education for All initiative and the Millennium Development Goals to achieve universal primary education of high quality the urgency to tackle high repetition rates increases. This research uses a unique panel data set from Senegalese primary schools in which the same students have been followed and tested over a period of five years to shed light on the impact of retention on subsequent achievement. We use a multi-level propensity score matching model that incorporates the development path of students with respect to their achievement and many other relevant student, teacher and school characteristics to infer the impact of retention. We find that in general retention has a negative effect on a student's achievement in subsequent years and for one specification that an initial positive effect vanishes over time. Given the overwhelmingly negative effect of retention even after three and four years we cast doubt on the effectiveness of high repetition rates.

1. Introduction

In developing countries high repetition rates are wide-spread. Repetition is the practice of schools to have students re-attend the same grade, and is often perceived as a measure to maintain or achieve educational quality. Several reasons can be found why students are not promoted to the next grade. One may be that some students are not as emotionally mature as their peers and are retained to have additional time to develop individually. Also, students may be required to obtain a certain level of knowledge to be promoted. If this level (e.g. measured by tests) is not achieved the student would have to repeat her grade. Furthermore, the practice of grade retention[1] may also act as an incentive or deterrent for students in order to encourage them to perform well and to put more effort in their studies.

In the broader context there are two initiatives formulating educational goals to which the issue of grade retention may be linked. The Education for All (EFA, cf. UNESCO (2000)) movement and the Millennium Development Goals (MDGs) that seek to achieve universal primary education of

[1] The terms "grade retention" and "grade repetition" are used synonymously throughout this paper

high quality. If repetition rates are very high, the educational system may quickly suffer from overcrowding if additional educational inputs and infrastructure is not provided. When enrolment rates stay constant there will be more students per class the higher the number of repeaters. With growing class size pupil-teacher ratios increase and the schools may not be able to manage high numbers of school enrolment. An education system that is overloaded as a result of high repetition rates therefore endangers the aims of the two initiatives.

While there is a heated debate about the efficiency of grade retention in developed countries, little research has been done with respect to developing countries. It is necessary to analyse developing countries separately as the educational framework there differs considerably. Repetition rates are considerably higher in developing countries. In the late nineties the average repetition rate per grade in the first four grades was about 20% in Francophone Sub-Saharan Africa, and about 15% in Anglophone Sub-Saharan Africa. The numbers did not change much in the following decade and were one to two percentage points lower in 2010. In Senegal repetition averaged roughly 10 per cent in 2000 and decreased to about 5 per cent in 2010[2]. Often these countries face problems that are virtually unknown in the developed world (e.g. low school attendance during harvesting season or large distance to the next school). Other problems, like the costs of retention, are similar although they impact differently (developing countries face different financial constraints). When repetition rates are extremely high, it is difficult to simply argue that all the repeating students are not mature enough or did not acquire the necessary knowledge to proceed to the next grade. Instead the question arises how effective these high repetition rates are and how much they actually benefit the individual student. We could think of at least two reasons why repetition may be seen as effective. The first being, that repetition enhances sub-sequent achievement of the student. The second possibility is a potential positive impact on the drop-out behaviour of the student leading to fewer drop-outs after retention. The latter relationship may also be linked to achievement as it is likely that well-performing students have a higher propensity to stay in school. In that sense, if retention impacts on achievement, it may also affect the drop-out behaviour.

Recent empirical evidence on the effects of grade retention in developed countries is ambiguous. For US high schools, Eide and Showalter (2001) report significant positive correlation between retention and drop out as well as significant negative correlation between repetition and wages after high school for their OLS estimates. When taking endogeneity into account by using an instrument for retention these effects disappear in their analysis. Jacob and Lefgren (2004) use a regression discontinuity design (RDD) to study the test based promotion policy in Chicago schools and show that repetition has a positive impact on test scores of third grade students, however no

[2] Source: UNESCO-UIS (2011) and own calculations

impact on the achievement of sixth grader. Furthermore, Jacob and Lefgren (2009), also using RDD and the same data find that the probability of drop out is not increased by retention for 6th graders, but for low-achieving 8th graders. The pattern of short term positive effects and the lack of an impact in the medium term are also reported by Alet (2010) for French students. The author uses an IV approach and reports that grade retention in grades 1 and 2 improves student achievement in the short run (grade 3) but loses its impact on test scores after several years, when students are in grade 6. Another study in France by Mahjoub (2008), using an IV approach as well as a matching approach finds a positive impact of grade repetition on student achievement and on her probability to graduate. Fertig (2004) also finds a positive effect of grade retention on educational outcomes (increased probability of obtaining a high schooling degree) in Germany, when tackling unobserved heterogeneity by instrumenting for grade retention.

The link between retention and labour market outcomes has only little been researched so far. One example is Eide and Showalter (2001) who do not find any significant results for their Chicago sample. Brodaty et al. (2010), however, find robust evidence for France, that delay, mostly stemming from grade retention decreases wages by as much as 9% and has a negative impact on employment overall.

Research on grade retention in developing countries is quite scarce and no less ambiguous than the literature in the developed world. The main problem being, that data from developing countries are hardly available and their quality is often inferior. Manacorda (2010) uses regression discontinuity to analyse Uruguayan public schools. He shows that automatic grade failure (when students miss more than 25 days of school) increases drop-out and impacts negatively on educational outcomes 4 to 5 years after failure. In contrast, Gomes-Neto and Hanushek's (1994) analysis of primary schools in Brazil's rural northeast finds that repetition has a positive impact on achievement scores in later grades. This last result, however, should be treated cautiously because of the small size and low quality of their sample.

In-depth studies on the impact of grade retention in Sub-Saharan Africa are almost inexistent, in particular regarding evidence identifying a clear causal relationship between retention and achievement and/or drop-out. Many studies regarding education in francophone sub-Saharan Africa use data from PASEC[3], an evaluation program of francophone education systems that has collected data from a number of francophone Sub-Saharan African countries.

André (2008) is a rare example for a study that deals with grade retention in Sub-Saharan Africa. Using PASEC data from primary schools in Senegal he uses an IV approach instrumenting grade retention by teachers' attitudes toward repetition and an RDD. He finds that repetition lowers

[3] Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN[3] (PASEC), where CONFEMEN stands for Conférence des ministres de l'Education nationale ayant le français en partage.

grade enrolment in the subsequent year. Bernard et al. (2005) analyse the five-year PASEC panels for Côte d'Ivoire and Senegal and find a strong negative relationship between previous grade retention and test scores in grade 5. Furthermore, they provide initial evidence that repetition increases the likelihood to drop out of school. In PASEC (2004) the panel for Senegal is used to examine the impact of repetition on subsequent achievement separately for different grades. In this analysis test scores of repeaters are compared to those of promoted students, but only for those test items that were common in both tests. The results suggest that throughout different grades repeaters perform worse on these tests than promoted students even if it is controlled for initial achievement and other variables. The study, however, merely relies on simple Ordinary Least Squares methods that are likely to suffer from omitted variable bias as unobserved student ability has not been accounted for.[4] As we have pointed out above retention, achievement and drop-out are closely interlinked. In this work, we focus on the essential relationship between repetition and achievement and leave the drop-out criterion for subsequent research work. Our analysis is an in-depth extension of the early evaluation in PASEC (2004) that tackles a number of issues regarding the content and overcomes important shortcomings with respect to the statistical methods. In contrast to PASEC (2004) we consider the development of students in terms of achievement several years after their retention (not just the first subsequent year) and do not confine ourselves to specific items in student test scores. More importantly, we target the problem of causal inference by including the development paths of students before retention which are likely to account for relevant (unobserved) covariates that have not been surveyed. By this methodology we strengthen the credibility of our statistical assumptions, in particular the Conditional Independence Assumption (CIA). The advantages towards earlier evaluations are detailed in the following sections. Section two elaborates on possible links between retention and sub-sequent student achievement. The third section gives details on the data used and data management. Section 4 explains the methodology we selected and reports the results of our analysis and section 5 concludes.

2. The link between retention and achievement

The achievement of a student is influenced by a number of variables that have been identified in previous literature, such as the socio-economic background of the student. Repetition is likely to have an effect on student performance via a number of channels. One greatly debated feature of retention in the literature from the developed countries is the fact that retained students are taken out of their social group which is seen as detrimental to the motivation of the student. The

[4] Recent PASEC reports on other francophone sub-Saharan African countries also use simple regression techniques for their cross-sectional data and find negative effects of retention on achievement (cf. e.g. PASEC (2009, 2010a, 2010b))

student will see her peers promoted to the next grade and therefore may perceive her academic failure as a personal one. The more often the student repeats the stronger the personal failure may be perceived. Moreover, it may well be that the student is mocked by her old and new peers who were constantly promoted so far and feel ashamed. In addition, the student will have to settle in her new class, starting anew to make friends and to get socially accepted by her peers. As the student will be overage compared to the new classmates the student will have to struggle harder to achieve acceptance from her peers and therefore suffer from motivational drawbacks (cf. Roderick 1994, 1995). These linkages are more likely to hold for the developed countries than for the developing world, as due to the high repetition rates per grade the individual repeater is less likely to be isolated or ashamed since quite a number of his old peers will repeat at the same time as her or in another year. Another reason for demotivation that is more specific to developing countries may lie in the awareness of the student of the additional costs she is causing by her grade repetition. If we perceive the schooling as a cost-benefit calculus/consideration as in (cf. Bedi and Marshall (2002) or Glick and Sahn (2010)), the cost of schooling will come about in opportunity costs as the student could work in the family business or labour market instead of attending school. Benefits for the student are human capital accumulation and thereby increased job market opportunities later in life. If a student is aware of the cost an additional year of schooling is causing for her or her family she may well be demotivated resulting in lower achievement or even lead to drop out. As students are able to work more the older and stronger they become (e.g. in agriculture) these opportunity costs magnify with increasing age.

Above we have named several reasons why students may have to repeat their grade (emotional immaturity, low achievement or incentive-building). As we will show later retention has an important impact on student achievement. Our main argument is that grade retention is one important determinant of student achievement among others that have been identified in previous literature. We will establish/identify the causal link between repetition and achievement and quantify its effect on achievement in the years after repetition.

In favour of grade retention it is sometimes argued that students who are retained have additional time to fill their knowledge gaps. We may consider a student who has been retained as a result of low test scores in important subjects and we therefore assume that there exist considerable knowledge gaps. If this student is promoted she may face difficulties to cope with the new contents as these will be more advanced and build on material taught in previous classes. The knowledge gap may therefore widen with each class. In contrast, even if the student had initial knowledge gaps, one could still argue that she will increase her effort and catch up with the course content because of her increased maturity or motivation. In the latter scenario, promotion would be the better choice. In any case when the student repeats there will be additional learning in this year – possibly to fill

knowledge gaps or to acquire further knowledge and the student should perform better relatively to her new peers than she did relative to her old peers in the grade retained. The differentiation between old and new peers is important when comparing the students. On the one hand we may analyse how the student develops compared to her new class mates (she will, however, have had one more year of schooling) or compared to her old classmates who have been promoted to the next grade (but have as many years of schooling as the repeating student). These two concepts will be detailed in section 4. Using the former idea of comparability and our initial rational that retention will give the student an additional year of schooling we posit our baseline hypothesis.

H1: Relative to their new peers, repeaters will do better than they did relative to their old peers in the previous year

The positive attitude toward repetition often found among teachers may be the result of the first type of comparison. The teacher may not compare a student who repeats to her old peers, but to her new peers and therefore possibly perceive retention as a useful measure where in fact it was not based on comparison to her new peers. (cf. CONFEMEN (2002))

An answer to the first hypothesis by itself will only give initial insight but it will help to shed light into the effects of repetition in a step-by-step procedure. Up to this point we have no indication yet whether possible greater absolute achievement in comparison to the new peers in the aftermath of retention is due to repeating a grade or a result of unnecessary repetition and therefore independent of the retention decision.

It is not easy to judge if filling knowledge gaps on the positive side or (de-)motivational aspects on the negative side have a stronger impact on subsequent achievement after retention. We tend to believe that the latter have a greater influence for two reasons. First, among repeaters there is apparently a considerable number of high and medium achieving students (PASEC 2004) whose retention is rather difficult to explain by knowledge gaps. Second, the results of recent PASEC reports of other francophone sub-Saharan African countries suggest that retention impacts rather negatively on subsequent achievement (e.g. PASEC (2009, 2010a, 2010b))

Let us assume that catching up and demotivation balance out on average. In that case a student who is retained will perform just as well academically if she had been promoted. If our consideration that demotivation is the stronger argument is true, we would expect that a retained student will perform worse in subsequent grades than if she had been promoted.

H2: Retained students acquire the same or lower level of knowledge and skills when compared to a matched control group of promoted students

Furthermore, when the student enters higher grades of primary school she will approach early adolescence and the difference between older and younger students may appear more obvious. These may come about in physical or emotional shape and emphasize the differences between older students (who have repeated in the past) and students of average age. Therefore the criterion of being overage will be the more pronounced the higher the grade. Hence, corresponding to the findings of Alet (2010) in France and Jacob and Lefgren (2004, ; p.235) in the US that retention will show a worse (or less positive) effect several years after retention than in the direct aftermath we hypothesize:

H3: The longer grade retention dates back the worse the performance of repeaters compared to a matched control group of promoted students

3. Data and Data Management

3.1 Variable Selection

The analysis of this paper uses exceptionally rich panel data from Senegalese primary schools surveyed during the period 1995-2000. The survey was conducted as part of PASEC in a cooperation between the Senegalese Ministry of Education and CONFEMEN starting with a cohort of roughly 2000 second graders in 1995 and following these students for five consecutive years unless they dropped out of school. The students were tested each year in both math and French. The tests were conducted at the end of each year in addition to a pre-test at the beginning of the second grade enabling the researchers to control for prior achievement. This survey was specifically designed to allow for a thorough analysis of repetition practices and their consequences in Senegalese primary schools. The data include a wide range of variables detailing the characteristics of the students and of their environment. These data come from questionnaires for students, teachers and directors giving insight into the socio-economic background of students, the schooling conditions and the characteristics of their teachers (cf. Bernard and Michaelowa (2005) and PASEC (2004)). 20 Students were selected at random in each of just under 100 schools to answer the questionnaire and participate in the tests leading to 1977 observations of which 1746 have a non-missing test score value. Since the number of dropouts is considerable the number of observations decreases from one survey wave to the next. In addition, some students were not at school on test days and account for some of the missing data in the test score variable. Therefore, there are only 1009 observations with a non-missing test score value left in the last year of the survey.

We classify the explanatory variables into three major categories: Student specific characteristics that include mainly the respective socio-economic status (see e.g. Dumas and Lambert (2011)) on the relationship between socio-economic background and schooling in Senegal) , class variables and teacher characteristics. We use five variables to identify the socio-economic status, namely the student's gender (BOY)[5], the work obligations outside school, the intake of meals, an index for property at home and an index for media availability at home. Work obligations outside school (WORK HOME) is a categorical variable comprising eight possible work fields: Cooking, cleaning, washing, agriculture, animal husbandry, dishwashing, childcare and commercial activities. This index shall account for the distraction from school activities and the reduced time available for homework and home study. The index for property at home (PROP HOME) details the wealth situation of the student's family comprising the possession of a car, a fridge, a flush toilet, electricity and a water tap. The index of media availability (MEDIA) contains information on the possession of a TV, radio and/or video device at home. The latter variable can be perceived as an additional wealth index. By the nature of these technical devices, it is further linked to education since consuming videos, radio and TV shows may improve a child's understanding in the spoken languages, drive curiosity or even be directly useful as educational transmitters when transferring useful knowledge. For this reason, it is separated from the previous property index. Finally, intake of meals (MEALS) measures if the student has regular breakfast, lunch and dinner. We consider this important as the link between nutrition and educational outcomes has been established in the literature (cf. Michaelowa (2000), Vermeersch and Kremer (2004) as well as Glewwe (2005)). Besides these student characteristics we include information on the personal schooling condition of the respective child. That is on the one hand the number of grades repeated up to the current grade (REPEAT PRIOR) of interest. We regard this to be important information since we want to measure the effect of a specific repetition decision that could in turn be influenced by prior repetition. What is more; with respect to personal schooling conditions we include current math test scores (TEST SCORES) which should be important in determining repetition.

We use three variables containing information on class or school characteristics[6]. One is the size of the class (CLASSIZE) (see e.g. Angrist and Lavy (1999),Case and Deaton (1999) and Hanushek (1998) on the impact of pupil-teacher ratios on student achievement) and the other two are the shares of available math and french books (BOOKM_CL, BOOKF_CL; see e.g. Fehrler et al. (2009) for a study in the context of sub-Saharan Africa) in the class of the student. Furthermore, we include a measure of urbanization that contains information on the urbanization of the school location differentiating between small village, big village, suburban area and town (CITYSIZE).

[5] In brackets the variable names are indicated as used in the estimations.

[6] As only one class per school has been sampled we use class and school characteristics as convenient.

The third group of variables, the teacher characteristics, is considered by variables for the gender of the teacher (MALE-T), the training she received (TRAINCUM-T) and the work experience she obtained (JOBEXP-T). The teacher training variable is a categorical one starting with a value of zero for no training up to a value of five for a training period of more than twelve months. Job experience contains teaching experience in number of years.

Table 1: Student-, Teacher- and Class-Specific Variables

Variable[1]	N	Mean	Std. Dev.	Minimum	Maximum
Student Characteristics					
REPEAT	5590	0.16	0.37	0	1
REPEAT PRIOR	7041	0.50	0.66	0	4
TEST SCORES 2	8676	-1.33	1.52	-6.21	3.98
TEST SCORES 3	7152	-0.39	1.50	-5.21	4.08
TEST SCORES 4	5705	0.79	1.51	-5.21	5.15
TEST SCORES 5	5827	1.45	1.49	-4.40	5.69
TEST SCORES 23	6612	0.84	1.04	-3.58	5.91
TEST SCORES 34	5435	1.07	1.19	-3.26	5.51
TEST SCORES 45	4635	0.66	1.22	-8.31	5.05
BOY	9695	0.55	0.50	0	1
PROP HOME 2	9696	3.16	2.36	0	7
PROP HOME 23	9695	0	0	0	0
PROP HOME 34	9695	0.37	1.30	-6	6
PROP HOME 45	9695	-0.20	1.12	-5	5
MEDIA 2	9695	1.50	0.91	0	3
MEDIA 23	9695	0	0	0	0
MEDIA 34	7065	0.17	0.79	-3	3
MEDIA 45	7035	-0.17	0.79	-3	3
WORK HOME 2	9780	2.64	1.91	0	8
WORK HOME 23	9780	0	0	0	0
WORK HOME 34	7150	0.84	1.82	-6	8
WORK HOME 45	7150	0	0	0	0
MEALS 2	9375	2.94	0.23	0.50	3
MEALS 23	9375	0	0	0	0
MEALS 34	5400	-0.79	0.59	-3	1.50
MEALS 45	5720	0	0	0	0
CITYSIZE 2	8975	3.10	1.15	1	4
CITYSIZE 23	8975	0	0	0	0
CITYSIZE 34	8975	0	0	0	0
CITYSIZE 45	8975	0	0	0	0
Teacher Characteristics					
MALE-T	5886	0.65	0.48	0	1
JOBEXP-T 2	9731	13.69	8.40	1	34
JOBEXP-T 23	8402	-0.46	8.71	-23	31

JOBEXP-T 34	5660	0.03	7.61	-31	18
TRAINCUM-T 2	9786	3.45	1.26	0	5
TRAINCUM-T 23	9332	0.11	1.65	-3	5
TRAINCUM-T 34	6990	0.03	1.10	-4	3
TRAINCUM-T 45	5280	0.04	0.89	-4	4
Class Characteristics					
BOOKF_CL 2	8920	0.50	0.28	0	1
BOOKF_CL 23	7340	0.23	0.26	-0.80	1
BOOKF_CL 34	6455	0.05	0.25	-0.88	1
BOOKF_CL 45	5110	-0.04	0.24	-1	0.67
BOOKM_CL 2	8920	0.30	0.27	0	1
BOOKM_CL 23	7275	0.04	0.27	-1	1
BOOKM_CL 34	6400	0.09	0.31	-1	1
BOOKM_CL 45	5275	-0.10	0.28	-1	0.86
CLASSIZE 2	9031	54.22	15.80	12	102
CLASSIZE 23	7432	-5.78	14.36	-42	61
CLASSIZE 34	5220	-0.29	13.83	-93	57

[1]Descriptive statistics of complete data set after partial imputation. Notation:
[Variable] 2 indicates statistics of respective variable in grade 2 – [Variable] 3/4/5 equivalently (provided for test scores only).
[Variable 23] indicates change of respective variable from grade 2 to grade 3 – [Variable] 34/45 equivalently.

3.2 Imputation

Virtually all the variables mentioned contain a considerable number of missing values. As we would like to achieve best possible inference from the data at hand we chose to impute the data wherever possible. We did not, however, resort to imputation commands of statistics packages but rather conducted these imputations “manually” to achieve best possible reliability. In some cases we had to decide which would be the best way of imputing data as was the case for the variable indicating work outside school. For this variable we imputed a missing value by the value of the precedent or following year as there was little variation in this variable. We believe these changes to be minor and necessary. They will provide us with more observations without greatly impacting on their reliability. As can be seen from the descriptive statistics in table 1 the panel is still imbalanced as we have not attempted to find a value for each missing since plausible values could not be found in a number of cases.

3.3 Item response theory

An excelling feature of our data is the availability of observations across time for the same students, specifically their achievement measured by test scores. Educational data with this property are almost inexistent in developing countries. In most studies – even for developed countries – where panel data are available, the comparability is severely restricted as the difficulty of the tests

varies with the different grades in which the students are tested. For this reason most studies referred to above confine their analyses to common items, i.e. questions that are identical in the tests of the grades that the researcher intends to compare. When using this methodology the problem arises that only a small part of the actual test is analysed leaving out potentially important information from other test questions. Furthermore, the more distant the analysed grades are the more difficult it will be to find a sufficient number of common items.

For this reason the opportunities of item response theory (IRT) were used where questions that are not identical in different tests are made comparable (see e.g. Hambleton et al. (1991)).[7] This methodology uses anchor items which are common items in the tests of different grades and serve to evaluate the difficulty of non-common items. The items are calibrated to a common scale so that comparability of different sets of items is ensured. The link between the ability of a student and the difficulty of an item can be illustrated by the item characteristic curve. The curve describes the probability of a correct answer for different ability levels. By this method the ability of a student should be measured independently from the difficulty of a test item and the difficulty of the test item independently from the ability of the student. As both are unknown they are jointly estimated by Warm (Maximum) Likelihood Estimation (WLE). For our panel data calibration was complex due to the length of the panel. WLE math scores, however, could be calculated allowing for a comparison of test scores across all tests and therefore across all years of our panel (cf. Warm (1989)).

4. Methodology and Results

This analysis is a substantial extension in content and methodology of a preliminary analysis by PASEC (2004) of the impact of repetition on achievement using these panel data. Apart from their descriptive analysis PASEC (2004) largely focused on the comparison of test scores among repeaters and non-repeaters between two consecutive years. Due to the comparability problems discussed above this analysis had to be reduced to the common items of the respective tests. As we have overcome this problem we can make use of more observations and broaden the analysis from just the consecutive year to more distant years in the future. This is an important aspect as we hypothesized that the longer repetition dates back the more the repeaters will underperform compared to a matched control group.

While PASEC (2004) has largely relied on ordinary least squares (OLS) methods we will make use of propensity score matching to infer the impact of grade retention on student achievement. We chose this method for a number of reasons. One advantage of this method over OLS is the lack of assumptions regarding the functional form, i.e. we will not necessarily need linearity for our model to work. More importantly, we wish to compare only existing observations. While OLS extrapolates into

[7] We would like to thank Christian Monseur, an expert in psychometrics, for calculation of the Rasch Scores.

regions where there are possibly no observations the matching algorithm ensures that results are based on the actual observations.

We would like to infer how students would have performed if they had not repeated. For this counterfactual analysis, it is necessary to find students with similar characteristics in order to compare them. The students are matched based on observable characteristics such as their background inter alia as discussed above. The method is based on the conditional independence assumption:

$$(1) \quad Y_{di} \perp\!\!\!\perp D_i | X$$

Where D takes the value 1 if the student i belongs to the treatment group and 0 if the student belongs to the control group. If this assumption is satisfied there is no selection problem once observable characteristics are controlled for. In other words, adjusting for student, school and teacher characteristics, the potential outcome Y does not depend on the participation status D. Accordingly all relevant characteristics would need to be included. As Rosenbaum and Rubin (1983) have shown we may use the probability of treatment instead where matching is based on a single number, the propensity score:

$$(2) \quad Y_{di} \perp\!\!\!\perp D_i | P(D = 1 \mid X)$$

Consequently, we may compare the probability of being treated instead of comparing all relevant characteristics X. The probability of treatment for each student is generally calculated by a logit or probit estimation. Considering our panel we have a multi-level data structure of at least two levels. Students are nested within classes. As students within the same class tend to be more alike we cannot guarantee independence between observations across classes. Therefore we opted for a multi-level logit function to estimate the treatment probabilities of each student. This estimation specifically takes into account the different levels and the correlation of student characteristics within a class.

Furthermore, we need to take into account the time dimension of our panel data. However, the assumption of independence between matched observations and the stable unit treatment value assumption (SUTVA) should not be violated. SUTVA states that the potential outcomes for one observation are independent of assignment to treatment of other observations. Therefore we do not match a student in 2nd grade to herself in 3rd grade or to herself in any other grade. By using this procedure we avoid the most obvious source of SUTVA violation. Nielsen and Sheffield (2009) argue that matching should be based on panels paying attention that control samples do not overlap. In

their example they match “Mali 1989-1998” with “Malawi 1991-2000” and “Mali 1979-1988” with “Niger 1981-1990”. Therefore they overcome the problem of possible dependence among country-year dyads.

In our case of repetition, however, it is not advisable to adopt this procedure in exactly the same manner as I am specifically interested in the impact of repetition in a certain grade, i.e. I will distinguish between students repeating 2nd grade, 3rd grade and so on. Consequently, our analysis will draw from the matching methodology used by Findley and Young (2011) who split panel data by year, match observations within these subsamples and reconstruct these partial data sets into a new panel data set. This procedure guarantees that the same observation is not matched based on different years. For our purposes it is not necessary to reconstruct the partial datasets. We will report on the results for each available primary school grade separately. For a specific example, consider students of third grade. A student who repeats in this grade will be matched based on her characteristics in this grade to a student who is promoted to fourth grade, based on that student’s characteristics in third grade. Then, the difference in test score between these two students one, two etc. years after the repetition of one of the students will be compared. In this way, the effect of repetition can be assessed over time. Unfortunately, however, this methodology does not make use of time trends before retention. Therefore, we will further adjust their method by including time trends.

When there is more than one time-observation before the repetition of a student it is useful to match panels with two or more years. For this purpose, we will adapt the methodology of Nielsen and Sheffield (2009) but restricting the matching algorithm to only match based on corresponding years, i.e. matching student i based on characteristics of 2nd and 3rd grade to student j in 2nd and 3rd grade in order to receive results for retention in specific grades.

There are two issues to consider: First, how many time periods before treatment are to be included in the matching? Second, based on how many explanatory variables can we match in the pre-treatment years? The first question is easy to answer with respect to our data. As we have only a short panel of five years we will include all pre-treatment years for this part of the analysis. That is, students repeating fifth grade for example will be matched based on four previous years (second through fifth grade). Students repeating third grade for another example will be matched based on two previous years (second and third grade). The second question is more difficult to answer. For the same example of a fifth grade repeater, we would need to include our 10 explanatory variables for four pre-treatment years resulting in 40 explanatory variables. It will therefore be quite difficult to achieve balance among all covariates which is crucial for the matching procedure to work. Furthermore, we wonder how much influence e.g. the class size of a student four years earlier would have on current retention of a student. Therefore, we chose to take into account the time dimension

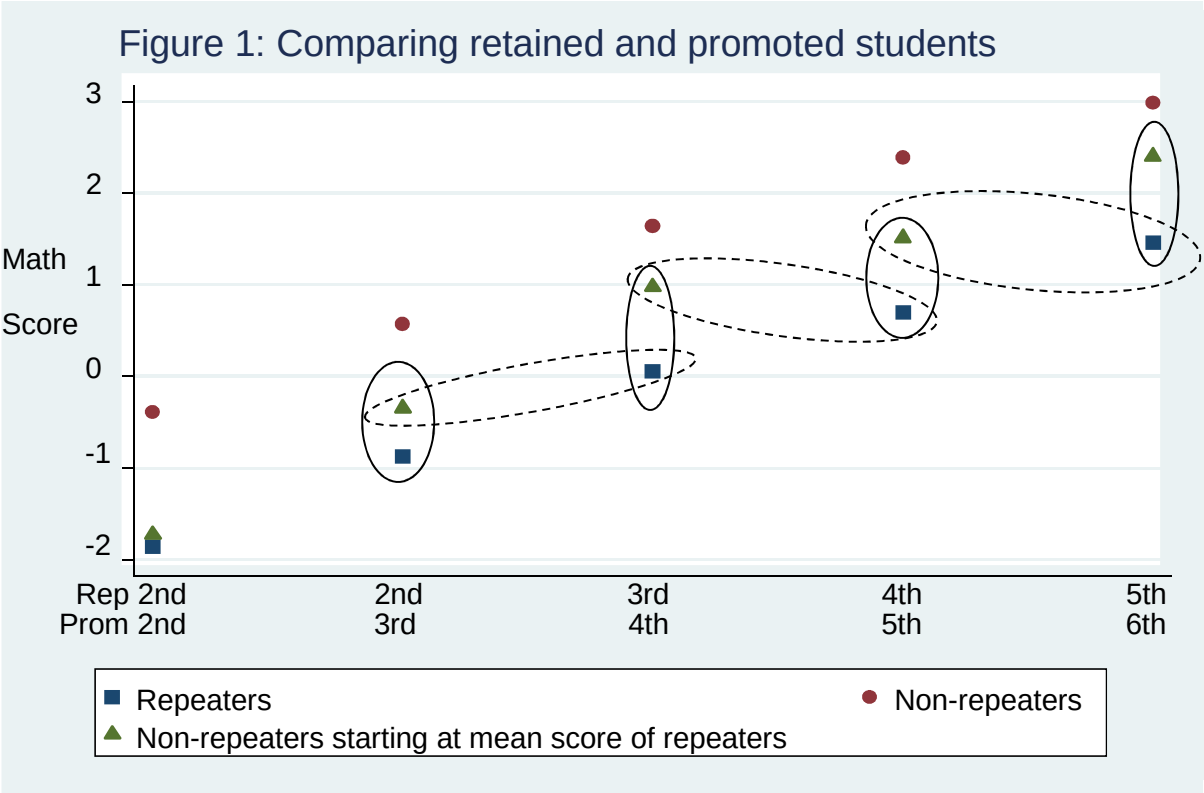
for the variable we judge most important for the retention decision: the test score. The test score is indeed the most significant variable in explaining retention which is evident from our propensity score estimation. Nevertheless, we will also report on our results with time trends of all explanatory variables. As we believe that the change in test scores from one year to the next is more relevant than the level of test score in each year, we included only the first year test score besides the changes from that year to the second, the change from second year test score to the third year one etc. The same procedure was used for all other covariates in the alternative specification that includes time trends for all variables.

It is important to note that that there are two differing concepts in comparing the treatment and control observations in the post treatment period. They reflect the varying views on the purpose that is attributed to grade retention. The first one (concept (a)) analyses how much a student should learn during one year of education. Therefore, if a student repeats in 3rd grade, we will compare her achievement one year later (he is still in third grade) to a non-repeating student who is in fourth grade by then. The idea is to analyze the changes in student achievement from one year to the next given repetition and non-repetition.

The second concept (concept (b)) contains the idea that a student should have a certain level of knowledge by a specific grade. Therefore, if a student repeats in 3rd grade, we will compare her achievement two years later (he then is in fourth grade) to a non-repeating student one year later who at that time is in fourth grade. When the latter is done we would like to find out if repeating enabled the student to reach the same level of knowledge in fourth grade as the non-repeating student. Note that in this case the repeating student had one more year of schooling to reach this level. The different possibilities of comparison are illustrated in figure 1.

It shows a second grade retention decision and follows the students till the end of the panel. Only those students have been selected who have not dropped out during the panel. The squares indicate students who repeated the second grade and therefore lag one grade behind the second grade non-repeating students which are shown by triangles. Concept (a) translates to a vertical comparison (solid circles) where the students had the same amount of schooling since the retention decision. Concept (b) can be applied by horizontal comparison (dashed circles) where a repeater when in third grade is compared to a non-repeater when in third grade and so on. As the mean test scores of non-repeaters is generally considerably higher (circular dots) we have selected those non-repeaters who roughly start at the mean of the promoted students in the first wave of the panel, i.e. they almost have the same starting point in terms of test scores. Referring to the vertical comparison we can clearly see that the repeating students build up a substantial negative gap in the post retention years towards their old peers. Accordingly, given the same amount of schooling the

promoted students will achieve higher scores than the retained students even though they had the same initial level.



The circles encompass the observations to be compared. The solid circles represent concept (a): The students compared have the same amount of schooling since the point of the repetition decision. The dashed circles represent concept (b): A 2nd grade repeater when in third grade is compared to a promoted student when in third grade et cetera. N=25 for repeaters and N=64 for non-repeaters starting at mean score of repeaters.

Following the rational of concept (b) we may now look at the horizontal comparisons. The retained students when in third grade have slightly higher test scores than the promoted ones had when in third grade. So far, if the policy is meant to guarantee a certain level of knowledge by a certain grade, retention could be seen as a successful means to achieve this aim. It is important to note that the retained student had one more year of schooling to obtain this level. Looking at fourth and fifth grade, however, the score of the retained students falls below that of the promoted student even though with an additional year of schooling as a result of retention. Therefore, we will have to be more cautious to support retention even in the case of a same-grade comparison in line with concept (b).[8]

Even though we have not controlled for other variables so far we can make an attempt to answer our first hypothesis based on figure 2. The figure shows the pathway in terms of test scores

[8] For variations of this figure with different starting points of promoted and retained students confer to figures 2 & 3 in the appendix. The pattern of all figures regardless of the starting values is roughly the same.

of students repeating second grade. Let us assume that the mean score in the second grade remained constant over time, i.e. the mean score of 2nd graders in the first year of the panel (which is known to us from our data) is the same as the mean score of 2nd graders in the following year. Using this assumption we can simply compare the mean score of repeaters when they attend 2nd grade the second time to their mean score when they attended 2nd grade in the first year of our panel. As the mean score of the repeaters has increased, we can state that the repeater will be better in relative terms when compared to her new 2nd grade peers than when compared to her old 2nd grade peers. This does not come as a surprise as the repeaters had an additional year of schooling and there was some improvement of scores in that year. The result of this simple and straightforward analysis corroborates the perception that teachers may believe retention to be useful as they can see the relative improvements of repeaters in their new class.

These concepts are also transferred to the matching analysis described above. One of the most important tasks in this kind of matching estimation is to make sure that there is balance between treatment and control group with respect to all covariates used in the estimation. That is, we will have to find observations that are indeed comparable and that do not differ significantly in their characteristics. We found that radius matching, which make use of a caliper leads to very good matches in most of our specifications. Caliper matching defines the maximum difference in propensity score between two observations that may be used for comparison. In contrast to nearest neighbour matching, caliper matching avoids bad matches when the best comparison observation is very different. We have defined this distance to be no more than 0.05. It is used in all but one specification where we used 0.03 to achieve better balance. For comparability we would have preferred to use the same specification throughout the estimations but we are convinced that balance should be the most important criterion that enables us to speak of a matched control group. One possible drawback compared to nearest neighbour matching – specifically in the case of few observations – is that using caliper may lead to fewer matches and may therefore result in a higher variance of the estimation. We follow the suggestion of Dehejia and Wahba (2002) who propose radius matching, that is to use all comparison observations within the caliper not just the nearest one to overcome this problem (cf. Caliendo and Kopeinig (2008)).

We will start our analysis by discarding the time dimension in our panel and split the data into four cross-sections covering the retention decision in second, third, fourth and fifth grade respectively (specification 1). This will serve as our baseline specification. Then we will report on the results of the specification including the time trends of the variable TEST SCORES (specification 2) and then discuss the results of the full specification including time trend for all covariates (specification 3). In all these specification we follow the logic of concept (a) and report on the outcome variable – the post-treatment test scores – one, two etc. years after the retention decision.

Table 2, partitioned into parts A to D, contains all results for concept (a) – sorted by time of outcome, specification and grade. For the first specification, that is for the cross sections we observe that the average treatment effect on the treated (ATT) is negative and significant (at the ten per cent level) for all grades if the test scores of the following year is used as outcome. The ATT ranges from -0.34 in grade 2 to -0.65 in grade 5. In this specification students are compared with respect to their socio-economic background, their achievement and the characteristics of teachers and the school in the year at the end of which the repetition decision is taken. According to our results, students who are equal or very similar with respect to these characteristics perform worse in the following year if they are retained than if they are promoted. In other words, even though retained students are equal to promoted ones regarding a large set of characteristics they will underperform as a result of retention. The size of this impact varies slightly at the grade level with more negative effects for higher grades. Looking at the distribution of test scores in the respective grades, the magnitude of most of the significant estimates lies between a quarter and a third of a standard deviation of test scores (cf. table 1 and 2). Moreover, note that in general the number of observations decreases in each grade due to attrition, that is drop out from school. Note the jump from 4^{th} to 5^{th} grade in most specification. It is due to the fact that we could not include CLASSIZE and JOBEXP-T in the estimation of the propensity score as we do not have data for these two variables for 5^{th} grade. These two variables have a number of missing values in earlier grades that led to exclusion of the respective observations. Therefore, the lack of these two variables in 5^{th} grade estimations results in a higher number of observations. The disadvantage being less reliable ATT estimates for 5^{th} grade as we believe these two variables to be important covariates that need to be controlled for.

We consider specifications 2 and 3 to be the more meaningful ones as earlier time periods are considered. Technically, this is equivalent to the inclusion of further covariates that strengthen comparability. We can, however, attribute at least two more advantages to the inclusion of time trends. First, as we include the time *trends*, that is the difference of test scores from one year to the next in specification 2 and of all covariates in specification 3 we have also included a development path of each student rather than merely a snapshot in time. Second, if it is perceived that there are possibly additional (unobserved) covariates that need to be included for better comparability across students than these covariates are likely to influence the pre-treatment covariates that we included and therefore be part of the set of covariates we controlled for. This idea is corroborated by the fact that matching is inter alia based on pre-treatment test scores in order to get an estimate for post-treatment test scores. That is, if an additional variable influences post-treatment test score than it is also likely to influence pre-treatment test scores and possibly also the development of pre-treatment test scores. Also note that there is no estimation for the 2^{nd} grade in specifications 2 and 3 as there

Table 2: Estimation of the impact of Retention on Achievement (ATT)[1]

A - Outcome After *ONE* Year		N[2]	ATT[5]	S.E.	T-stat
Specification 1	2nd grade	136 (126)	**-0.34**	0.20	-1.72
	3rd grade	138 (137)	**-0.54**	0.16	-3.31
	4th grade	52 (50)	**-0.56**	0.24	-2.32
	5th grade	143 (122)	**-0.65**[4]	0.25	-2.62
Specification 2	2nd grade	-	-	-	-
	3rd grade	129 (123)	**-0.50**	0.18	-2.82
	4th grade	40 (38)	**-0.46**	0.27	-1.72
	5th grade	138 (105)	-0.28[4]	0.24	-1.19
Specification 3	2nd grade	-	-	-	-
	3rd grade	119 (112)	**-0.48**	0.18	-2.67
	4th grade	42 (39)	**-0.46**	0.25	-1.83
	5th grade	96 (59)	**-0.67**	0.23	-2.95
B - Outcome After *TWO* Years		**N**	**ATT**	**S.E.**	**T-stat**
Specification 1	2nd grade	85 (75)	-0.18	0.23	-0.77
	3rd grade	98 (92)	-0.04[3]	0.17	-0.21
	4th grade	32 (31)	**-0.59**	0.23	-2.52
Specification 2	2nd grade	-	-	-	-
	3rd grade	93 (88)	-0.13	0.20	-0.63
	4th grade	25 (22)	**-0.48**	0.27	-1.82
Specification 3	2nd grade	-	-	-	-
	3rd grade	80 (73)	-0.02	0.20	-0.09
	4th grade	20 (19)	-0.45	0.29	-1.54
C - Outcome After *THREE* Years		**N**	**ATT**	**S.E.**	**T-stat**
Specification 1	2nd grade	79 (67)	-0.35	0.22	-1.61
	3rd grade	62 (60)	-0.05	0.20	-0.22
Specification 2	2nd grade	-	-	-	-
	3rd grade	60 (57)	**-0.20**	0.23	-0.89
Specification 3	2nd grade	-	-	-	-
	3rd grade	48 (47)	-0.28	0.22	-1.26
D - Outcome After *FOUR* Years		**N**	**ATT**	**S.E.**	**T-stat**
Specification 1	2nd grade	65 (56)	**-0.38**	0.19	-1.99
Specification 2	-	-	-	-	-
Specification 3	-	-	-	-	-

[1] ATT computed using STATA's PSmatch 2 (Leuwen and Sianesi 2003) reprogrammed using the runmlwin command (Leckie and Charlton 2011) to estimate the mutli-level propensity score in MLwiN (Rasbash et al. 2009).
[2] Number of observations on support in brackets
[3] Caliper of 0.03 was used to achieve balance
[4] Balance after matching could not be achieved for one or more covariates
[5] Bold numbers indicate significance at 10% level

are no earlier time periods to be included.[9] These two variations do not differ much from the first specification. All the estimates of the ATTs are negative and – with one exception – slightly smaller than in specification 1. The estimate of the fifth grade ATT of specification 2 turns insignificant though. The first impression of a negative effect of retention is thus corroborated in the more encompassing specifications 2 and 3. In other words, given that students are equal or very similar based on the characteristics stated above in the first year of the panel as well as based on the development of these characteristics over time until the point of the repetition decision, then those students who repeat perform worse in the following year than those who do not repeat. These results give us insight into the dynamics of achievement in the year immediately following the repetition decision. This could be a special year as the repeating student is now for the first time in her new class with her new peers. Therefore, it is of special interest what the effect of retention will be after two and more years when the student has had time to get accustomed to her new peers and her new environment. Parts B through D of Table 2 contain the ATT estimates for student achievement more than a year after the retention decision. All estimates remain negative but render insignificant in many cases. Looking at these results for the outcome after more than a year we infer that repetition has either a clear negative effect or that no effect can be attributed to it. It seems clear, however, that following this particular type of comparison based on concept (a), there is no positive effect of repetition visible in our data set. For a closer look, we have rearranged the results in an alternative table for clarity.

Table 3: Outcome by Year and Specification

Outcome After		***One* Year**	***Two* Years**	***Three* Years**	***Four* Years**
Specification 1	2nd grade	**-0.34**	-0.18	-0.35	**-0.38**
	3rd grade	**-0.54**	-0.04	-0.05	
	4th grade	**-0.56**	**-0.59**		
	5th grade	**-0.65**			
Specification 2	2nd grade	-	-	-	
	3rd grade	**-0.50**	-0.13	**-0.20**	
	4th grade	**-0.46**	**-0.48**		
	5th grade	-0.28			
Specification 3	2nd grade	-	-	-	
	3rd grade	**-0.48**	-0.02	-0.28	
	4th grade	**-0.46**	-0.45		
	5th grade	**-0.67**			

See notes of table 2

[9] In Fact, specification 2 would be possible for 2nd grade as comparable scores for a test conducted at the beginning of 2nd grade have been computed. We are, however, not convinced by the quality of these scores and prefer to focus on those obtained from the tests at the end of each grade.

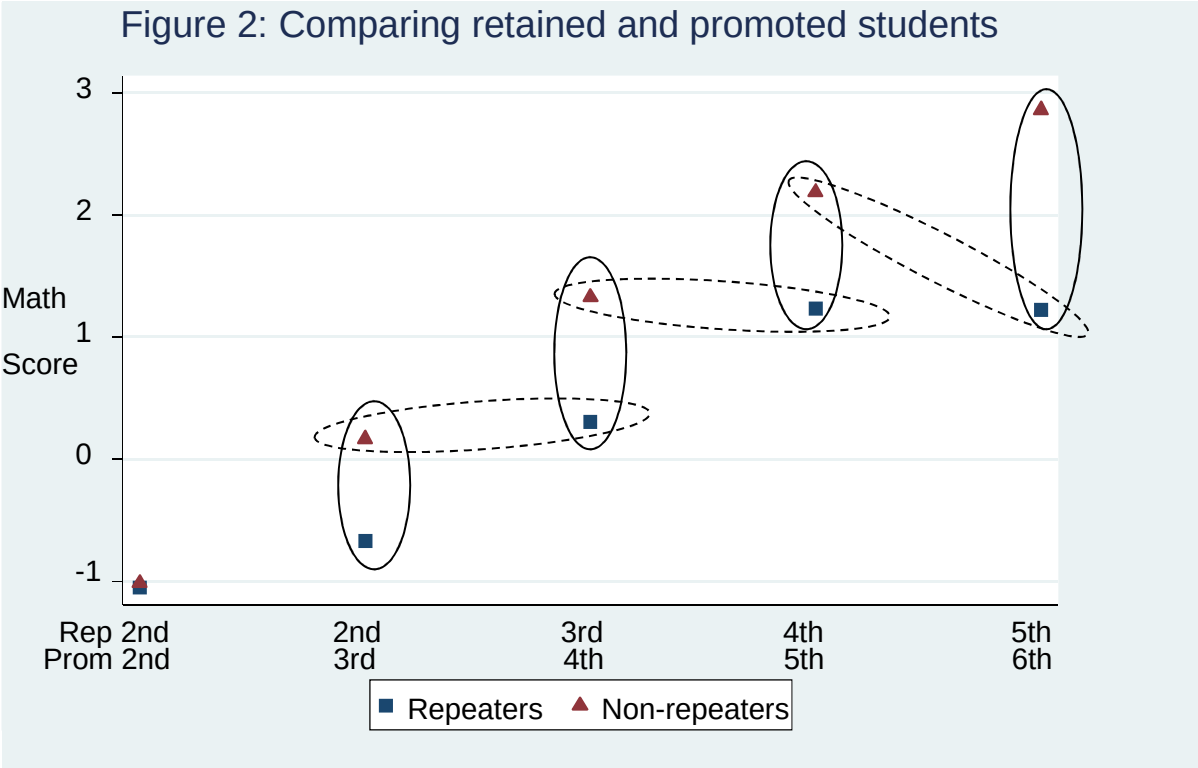
Table 3 contains the ATTs that we saw before ordered by outcome variable focussing on the estimates and leaving out the additional information of table 2 which makes a comparison of results over time more straightforward. Neglecting the lack of significance in a few cases we notice that after two years the size of the effect shrinks in almost all specifications of all grades. For grade 2 in specification 1 the absolute value then increases again in the third year (-0.35) (but remaining insignificant) and becomes even more negative (-0.38) after four years where it re-appears statistically significant. Similarly in third grade of specifications 2 and 3 the absolute value of the ATT re-increases after three years. In the case of specification 2 it amounts to -0.20 and re-appears significant as well. Fourth grade seems to be an exception where the absolute value of the ATT after two years rather increases in the first two specifications. In general, it seems that after an initial worsening impact of the repetition, the student recovers just to fall back again in years further away. Therefore,

the results suggest that unlike in the studies of Alet (2010) and Jacob and Lefgren (2004, ; p.235) achievement score further away are not continuously worse than in the direct aftermath. How may this pattern be explained? It is straightforward to imagine that the grade of repetition is the most difficult one for the student as she will enter a new social group and will have to handle the demotivation stemming from retention and the fact of studying for the same subjects again. What happens thereafter is less clear. Possibly, the student has managed to accept the new environment and has made friends. The worsening effect after more years may be explained by the reasons stated in the theoretical part of this analysis: Approaching early adolescence age differences may become more evident in physical and emotional shape widening the gap between retained and promoted students. These differences could lead to mockery by other students or de-motivate the overage student if she believes that she should be in a class of even-aged students. As students in fourth grade are nearer to early adolescence the gap is already visible in the first year after retention. This rational could explain why we do not see a recovery after fourth grade retention in specifications 1 and 2 (ATTs of -0.59 and -0.48). Based on these estimations we may now attempt to answer hypotheses 2 and 3. As for hypothesis 2 we can state that grade retention indeed has a negative impact on achievement or in some cases no effect. Referring to the time dimension of hypothesis 3 we cannot offer a clear statement. It seems that the initially very negative effect of repetition in the direct aftermath of the retention decision is mitigated when retention dates back two year just to be reinforced after three and four years. The effect after more years is not in all cases worse than in the initial period after retention. Conversely, given the negative signs for all coefficients, a positive effect cannot be attributed to retention in any of the specifications and grades.

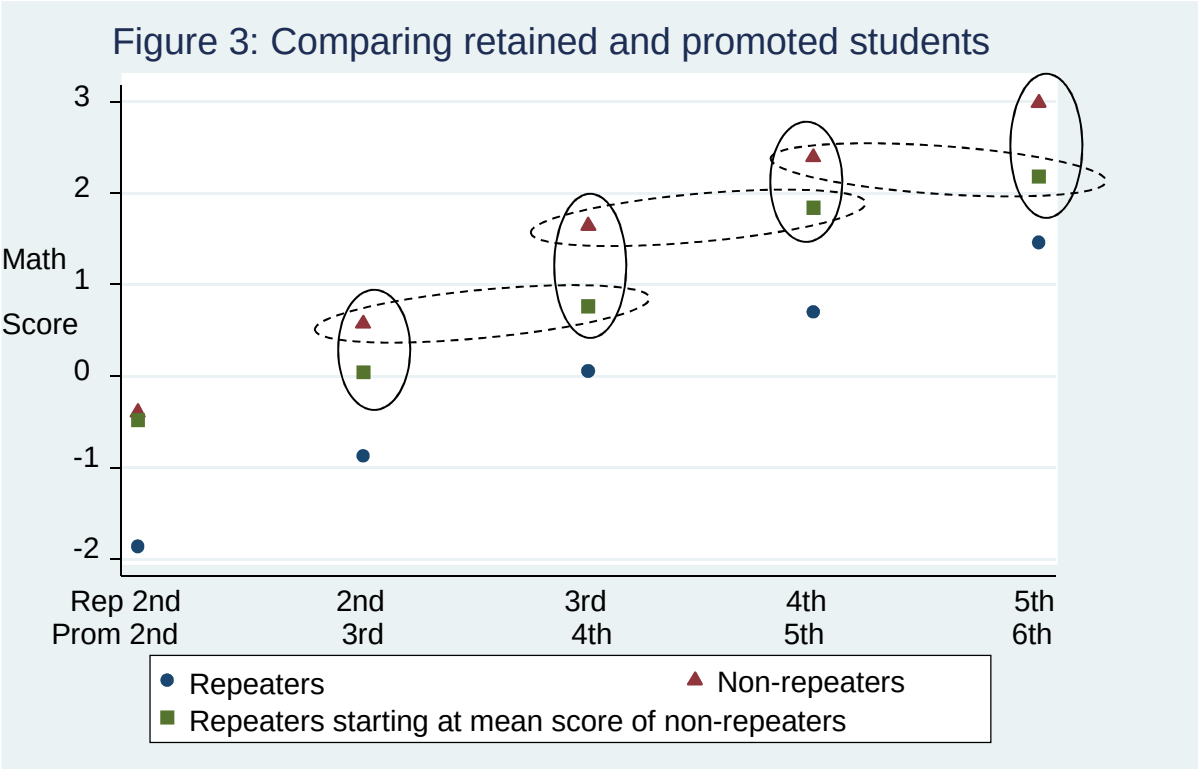
5. Conclusion

In this analysis we have researched the dynamics of retention over time for a variety of grades and specifications. We have used a unique dataset that follows a large number of Senegalese primary students for five consecutive years starting in 2nd grade. It includes a large variety of relevant variables on the socio-economic background of the students as well as on their teachers and schools. In accordance with the literature on repetition in francophone Africa we believe that retention renders higher cost for the education system of Senegal than it can offer as benefits. We have used descriptive graphic evaluation on the one hand and inferential statistical analysis on the other for the analysis. The statistical part consisted of propensity score matching that calculates the propensity score by a multi-level logit function including the development of student test scores over time and variables detailing student, teacher and school characteristics as covariates. Based on the results of these methods we observed that either a possible initial positive effect of retention turns negative after few years (descriptive graphic evaluation) or that retention does not seem to have any positive effect on student achievement at all (matching results). Against the backdrop of the aims of the EFA initiative and the MDGs these results appear highly relevant. The goal of universal primary education of high quality is ambitious and will be even more so if the education system gets overcrowded because of repetition. For the first time, we have shown in the context of francophone sub-Saharan Africa that there is little to expect from a retention policy that promotes high repetition rates with respect to educational quality (measured as student achievement) even beyond the initial negative effect of retention. This example from Senegalese primary schools is also highly relevant for other francophone sub-Saharan African countries where repetition rates remain high.

Annex

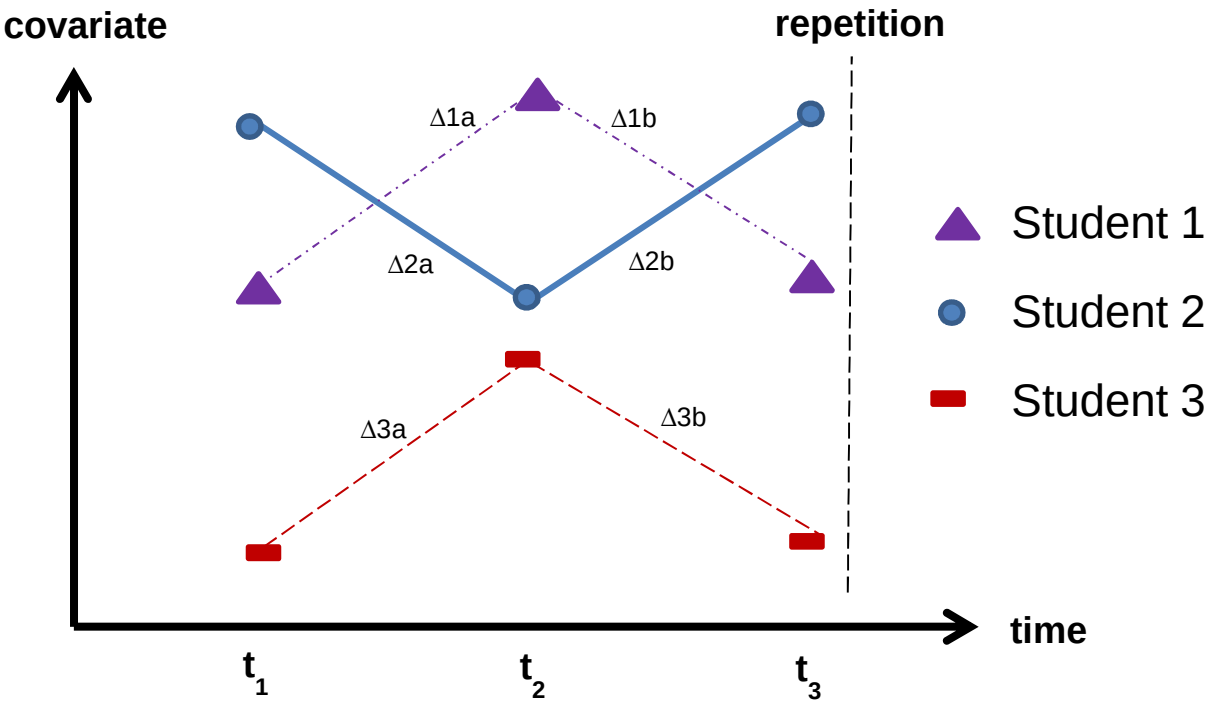


The circles encompass the observations to be compared. The solid circles represent concept (a): The students compared have the same amount of schooling since the point of the repetition decision. The dashed circles represent concept (b): A 2nd grade repeater when in third grade is compared to a promoted student when in third grade et cetera. N=9 for repeaters and N=93 for non-repeaters.



The circles encompass the observations to be compared. The solid circles represent concept (a): The students compared have the same amount of schooling since the point of the repetition decision. The dashed circles represent concept (b): A 2nd grade repeater when in third grade is compared to a promoted student when in third grade et cetera. N=7 for repeaters starting at mean score of non-repeaters and N=333 for non-repeaters.

Figure 4: Allocation of matches



According to specification 2 and 3 students are matched based on the value of the covariate(s) in t1 and based on the changes from one year to the next (Δ1a, Δ1b etc.)
In this example student 1 is matched to student 3 even though each data point of student 2 is closer to that of student 1. The idea is that the development path (Δ) is the most important criterion.

References

Alet, Elodie. 2010. "Is grade repetition a second chance?" In. Tolouse: Tolouse School of Economics.

André, Pierre. 2008. *The effect of grade repetition on school dropout*

An identification based on the differences between teachers.

Angrist, Joshua D., and Victor Lavy. 1999. Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement. *Quarterly Journal of Economics* 114 (2): 533-575.

Bedi, Arjun S., and Jeffery H. Marshall. 2002. Primary school attendance in Honduras. *Journal of Development Economics* 69: 129-153.

Bernard, Jean-Marc, and Katharina Michaelowa. 2005. "Managing the Impact of PASEC Projects in Francophone Sub-Saharan Africa." In *Cross-National Studies of the quality of Education: Planning their design and managing their impact*, eds. Kenneth Ross and Ilona Jürgen-Genevois. Paris: IIEP/UNESCO.

Bernard, Jean-Marc, Odile Simon, and Katia Vianou. 2005. "Le redoublement: mirage de l'école africaine?" In. Dakar: PASEC/CONFEMEN.

Brodaty, Thomas O., Gary-Bobo Robert J., and Ana Prieto. 2010. "Does Speed Signal Ability? The Impact of Grade Retention on Wages." In.

Caliendo, Marco, and Sabine Kopeinig. 2008. SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING. *Journal of Economic Surveys* 22 (1): 31-72.

Case, Anne, and Angus Deaton. 1999. School Inputs and Educational Outcomes in South Africa. *Quarterly Journal of Economics* 114 (3): 1047–1084.

CONFEMEN. 2002. "Les facteurs de l'efficacité dans l'enseignement primaire : données et résultats sur cinq pays d'Afrique et de l'Océan Indien." In.

Dehejia, Rajeev H., and Sadek Wahba. 2002. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics* 84 (1): 151-161.

Dumas, Christelle, and Sylvie Lambert. 2011. Educational Achievement and Socio-economic Background: Causality and Mechanisms in Senegal. *Journal of African Economies* 20 (1): 1-26.

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

Fehrler, Sebastian, Katharina Michaelowa, and Annika Wechtler. 2009. The Effectiveness of Inputs in Primary Education: Insights from Recent Student Surveys for Sub-Saharan Africa. *Journal of Development Studies* 45 (9): 1545-1578.

Fertig, Michael. 2004. *Shot across the Bow, Stigma or Selection? The Effect of Repeating a Class on Educational Attainment*.

Findley, Michael G., and Joseph K. Young. 2011. Terrorism, Democracy, and Credible Commitments1. *International Studies Quarterly* 55 (2): 357-378.

Glewwe, Paul. 2005. The impact of child health and nutrition on education in developing countries: Theory, econometric issues, and recent empirical evidence. *Food & Nutrition Bulletin* 26 (16): 235-250.

Glick, Peter, and David E. Sahn. 2010. Early Academic Performance, Grade Repetition, and School Attainment in Senegal: A Panel Data Analysis. *The World Bank Economic Review* 24 (1): 93-120.

Gomes-Neto, João Batista, and Eric A. Hanushek. 1994. Causes and Consequences of Grade Repetition: Evidence from Brazil. *Economic Development and Cultural Change* 43 (1): 117-148.

Hambleton, R. K., H. Swaminathan, and H. J. Rogers. 1991. *Fundamentals of Item Response Theory*. SAGE Publications.

Hanushek, Eric A. 1998. "The Evidence on Class Size." In *Occasional Paper*.

Jacob, Brian A., and Lars Lefgren. 2004. Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *The Review of Economics and Statistics* 86 (1): 226-244.

———. 2009. The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics* 1 (3): 33-58.

Mahjoub, Mohamed-Badrane. 2008. "The Treatment Effect of Grade Repetitions." In. Paris: Paris School of Economics.

Manacorda, Marco. 2010. The Cost of Grade Retention. *Review of Economics and Statistics*.

Michaelowa, Katharina. 2000. Dépenses d'éducation, qualité de l'éducation et pauvreté : L'exemple de cinq pays d'Afrique francophone. *OECD Development Centre Working Papers* 157.

Nielsen, Rich, and John Sheffield. 2009. "Matching with Time-Series Cross-Sectional Data." Presented at the Prepared for Polmeth XXVI, Yale University.

PASEC. 2004. *Le redoublement: Pratiques et conséquences dans l'enseignement primaire au Sénégal*. Dakar: CONFEMEN.

———. 2009. *Les Apprentissages Scolaires au Burkina Faso: Les Effets du Contexte, Les Facteurs pour Agir*. CONFEMEN.

———. 2010a. *Enseignement Primaire: Quels Défis pour une Education de Qualité en 2015?* Dakar: CONFEMEN.

———. 2010b. *Diagnostic et Préconisations pour une Scolarisation Universelle de Qualité*. CONFEMEN.

Roderick, Melissa. 1994. Grade Retention and School Dropout: Investigating the Association. *American Educational Research Journal* 31 (4): 729-759.

———. 1995. Grade Retention and School Dropout: Policy Debate and Research Questions. *Phi Delta Kappa Research Bulletin* 15: 1-6.

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

UNESCO. 2000. "The Dakar Framework for Action - Education for All: Meeting our Collective Commitments." In. Dakar.

Vermeersch, Christel, and Michael Kremer. 2004. "School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation." In: World Bank Policy Research Working Paper No. 3523.

Warm, Thomas. 1989. Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54 (3): 427-450.

Statistics Packages

Leckie, G. and Charlton, C. (2011). runmlwin: Stata module for fitting multilevel models in the MLwiN software package. Centre for Multilevel Modelling, University of Bristol.

Leuven, E. and B. Sianesi (2003). "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing". http://ideas.repec.org/c/boc/bocode/s432001.html. Version 4.0.4

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2009). MLwiN Version 2.1. Centre for Multilevel Modelling, University of Bristol.