CPB Discussion Paper

No 144

The effect of accountability policies in primary education in Amsterdam

Victoria Chorny, Dinand Webbink

Centraal Planbureau
Van Stolkweg 14
Postbus 80510
2508GM Den Haag

Telefoon (070) 3383380
Telefax (070) 3383350
Internet www.cpb.nl

ISBN 978-90-5833-446-6

Abstract in English

In 1995, the municipality of Amsterdam introduced accountability policies for schools in primary education. Population statistics show a large increase of test scores in the decade after the introduction of the new urban policies. This paper assesses this increase in test scores by analyzing data of a large sample of schools including scores on the published test and scores on similar independently taken tests that are not published. Difference-in-differences estimates show that after the introduction of the accountability policies, test scores for both tests taken in grade 8 increased substantially more in Amsterdam than in the rest of the country and more than in a sample of Low SES students. Approximately 60 percent of the increase of the published test scores can be attributed to an increase in general skills and 40 percent to an increase in test-specific skills. Test scores of pupils in lower grades also improved in Amsterdam. We do not find evidence for strategic behavior of schools. Although part of the gains in test scores might be test-specific, the accountability policies in Amsterdam seem to have succeeded in raising educational achievements in primary schools.

Key words: accountability policy, educational performance, primary education
JEL code: I20, I21, R00

Abstract in Dutch

Sinds 1995 is het gemeentelijk onderwijsbeleid in Amsterdam expliciet gericht op het verbeteren van de leerprestaties gemeten met de CITO-toets. Algemene statistieken laten een sterke stijging zien van de scores op de CITO-toets in Amsterdam. Deze studie onderzoekt deze stijging door Amsterdam te vergelijken met de rest van Nederland en met een steekproef van leerlingen met een lagere sociaaleconomische achtergrond. De studie gebruikt gegevens van het PRIMA-onderzoek dat zowel resultaten bevat van de CITO-toets als resultaten van toetsen voor taal en rekenen die zijn afgenomen binnen het PRIMA-project. De prestaties in Amsterdam zijn sterk verbeterd ten opzichte van die in de vergelijkingsgroepen, zowel op de CITO-toets als op de taal- en rekentoets in PRIMA. Ongeveer 60 % van de totale vooruitgang op de CITO-toets kan toegeschreven worden aan een algemene verbetering van de leervaardigheden en ongeveer 40 % aan specifieke vaardigheden voor het maken van de CITO-test. De prestaties van kinderen in Amsterdam in lagere groepen zijn ook verbeterd. Er is geen bewijs gevonden voor strategisch gedrag van scholen zoals het uitsluiten van zwakke leerlingen van de toets. Hoewel een deel van de vooruitgang in Amsterdam mogelijk bestaat uit een verbetering van specifieke vaardigheden voor het maken van de CITO-toets, lijkt het beleid in Amsterdam wel degelijk geresulteerd te hebben in een verbetering van leerprestaties in het basisonderwijs.

Steekwoorden: Gemeentelijk onderwijsbeleid, opbrengst gericht, CITO-toets

Contents

1 Introduction

A remarkable case of urban educational policy can be found in the city of Amsterdam. Within a decade, schools in Amsterdam have increased test scores on average with more than 0.5 standard deviation and surpassed the national average.[1] This strong increase in test scores happened after the introduction of a municipal education policy in 1995. This policy set targets for the level of test scores and focused on accountability of schools. Participation in a standardized national test (the CITO-test) was made compulsory for all primary schools in Amsterdam. The results of individual schools were published and schools received additional resources conditional on the improvements in performance. This paper takes a closer look at the increase in test scores in primary schools in Amsterdam.

Although the population statistics suggest that the educational policies of the city of Amsterdam have been a great success, the recent economic literature on school accountability policies suggests that caution is needed. Several recent studies show that school accountability policies can increase test scores (Hanushek & Raymond, 2005, Jacob, 2005, Dee & Jacob, 2009). However, strategic behavior of schools seems to be a fact of life. For instance, schools have raised test scores by classifying students as disabled (Figlio and Getzler, 2006) or increasing suspensions of low-performing students during the testing window (Figlio, 2006). Other strategic reactions that have been found are teacher cheating (Jacob and Levitt, 2003) and adding additional calories to the school menu on testing days (Figlio and Winicki, 2005). Jacob (2005) found that the test-based accountability policy in Chicago improved test scores but part of the improvement was related to an increase in test-specific skills. Figlio and Rouse (2006) show that the accountability policies in Florida only improved test scores in the high stakes grade.

Our paper contributes to the literature on the effects of accountability policies in education. Whereas most previous studies focused on the US by exploiting variation between states, we provide evidence for a European country and use variation within the same education system. The accountability system studied in this paper differs from the typical high stakes testing systems in the US with relatively strong incentives for low performing schools and students. The urban policy in Amsterdam set school-specific short and long term targets depending on the socioeconomic composition of the school population and the performance in previous tests. As such, the accountability policy in Amsterdam aimed at all schools.

We assess the increase in test scores in Amsterdam by analyzing data from a large country-wide sample of schools. A special feature of these data is that it includes both the published test scores and scores on independently taken tests that were not published. The data include two samples: a representative sample for the whole country and a ‘Low SES sample’ with an oversampling of pupils with a lower socioeconomic background. The data enable us to compare

[1] The average test scores on the standardized national test increased from 529.4 in 1996 to 536.6 in 2005 (one standard deviation is equal to 10 points). Amsterdam includes approximately 200 primary schools.

the change in performance of pupils in Amsterdam before and after the introduction of the new policies with the change in performance in these two samples. We estimate difference-in-differences models for both the published and not published test scores. The estimates for the published test scores are informative about the robustness of the gains in test scores for changes in the composition of the student populations that might have occurred since the introduction of the new urban policies. The estimates for the test scores that have not been published learn whether the impact of the Amsterdam policies can be explained by an increase in test specific skills ('teaching to the test'). In addition, our data include test scores of pupils in grade 2, 4 and 6 (the published nationwide test is only taken in grade 8). An increase in test scores of the unpublished test, especially for pupils in grade 2, 4 or 6, would indicate that the new urban policies really improved learning skills, and not only the ability to perform well on the published test taken in grade 8. A second important feature of the data is that it enables us to analyze non-participation on the published test. The new urban policies might have induced schools to exclude weak students from taking the test. We are able to compare the exclusion decisions of schools in Amsterdam with the decisions of schools in the rest of the country.

We find that the scores on the published test, which is only taken in grade 8, increased 0.5 standard deviation more in Amsterdam than in the rest of the country and 0.4 standard deviation more than in the Low SES sample. Especially the performance of 'regular Dutch pupils' improved but we also observe an improvement for ethnic minority students. We also find that test scores on the unpublished tests increased in Amsterdam more than in the other two samples after the introduction of the policy. The improvement of performance is found for pupils in grade 8 but also for pupils in earlier grades. However, the size of the improvement is smaller for the unpublished test and smaller for pupils in earlier grades. Approximately 60 percent of the increase on the published test in grade 8 can be attributed to an increase in general skills and 40 percent to an increase in test-specific skills. In addition, we do not find evidence that the increase in test scores is driven by the exclusion of pupils. Our overall assessment of the accountability policies in Amsterdam is positive. Although part of the gains in test scores might be test-specific the policies seem to have led to a substantial improvement of general skills of pupils in Amsterdam.

2 The accountability policies in Amsterdam

In the Dutch education system, the municipal education authorities are responsible for compliance with the compulsory education law, school buildings, the administration of public schools and education policies for disadvantaged groups. At the end of primary education, at the age of 12, pupils are tracked into different levels of secondary education. In the early nineties local education authorities in Amsterdam were concerned about the transition of pupils from primary to secondary education. Figures on drop out in the school year 1992/1993 showed that compared to the rest of the country drop out in Amsterdam was dramatically higher in all levels and grades of secondary education (Municipality of Amsterdam, 1994). The absence of admission rules for different tracks of secondary education was considered as an important factor for these high dropout rates. In the school year 1994/95, a so-called 'School Choice Procedure' was introduced which included various steps to improve the transition from primary to secondary education (Visser, 2003). One of these steps was that pupils should take a test at the end of primary education. The outcome of this test should be used as a second indicator of the ability of the pupil. The first indicator was the advice of the principal of the primary school about the secondary track. This 'School Choice Procedure', which made the use of standardized tests at the end of primary education acceptable for schools, preceded and became part of the broader municipal education policies that were introduced since 1995. The municipal accountability policies consisted of four-year plans called 'Towards Better Results' (Naar betere resultaten). The first four-year plan, focused on the period 1995-1998, can be seen as an initial phase in which ideas were introduced such as setting targets and monitoring of performance. The second plan, which focused on the period 1998-2002, was much more explicit in setting targets, using incentives and holding schools accountable for student performance. With the introduction of the second plan, the accountability components became the central elements of the urban educational policy. It has been suggested that the responsible alderman was inspired by ideas about public sector management through output steering from the new central government (Visser, 2003).

For the period 1995-1998, the local education authorities in Amsterdam reached an agreement with schools which was laid down in the plan 'Towards better results' (Naar betere resultaten). This plan included a general target and several activities. The general target was to raise the performance of pupils towards the national average. The first step in the new plan was the measurement of the performance of primary schools which suggests that the formulation of the target was not based on a quantitative assessment of performance. Visser (2003) notes that there was a general feeling that primary schools in Amsterdam were underperforming. The dramatic dropout rates in secondary education might have contributed to this feeling. It was agreed that from the school year 1995-1996 onwards the performance of pupils at the end of primary education would be measured in such a way that schools in Amsterdam could be compared with each other and with the rest of the country. At the start in 1996, 178 out of 207

schools in Amsterdam participated in the nationwide CITO test. Only the so-called Montessori schools refused to participate because of some basic objections against the CITO-test. However, these schools started participating in the school year 1997-1998. In 1996 schools in Amsterdam scored on average 5 points below the national average on the CITO test. In 1997, nearly the same number of schools in Amsterdam participated (177). The average score in Amsterdam was 4.6 points below the national average. The average scores of Amsterdam and the rest of the country were published in the yearly report over 1997 by the statistics agency of Amsterdam (Statistics Amsterdam, 1998). To our knowledge, this is the first publication in which the performance of schools in Amsterdam has been compared with the performance of schools in the rest of the country. Note that this is a comparison of the average score of Amsterdam with the average score in the rest of the country without any reference to the performance of individual schools. Another important component of the plan was the introduction of the use of test score ranges (band widths) for the assignment of pupils to different tracks in secondary education in the school year 1996-1997. These band widths created more transparency in the use of admission rules for different tracks of secondary education.

The municipality of Amsterdam reached a second agreement with schools and city districts for the period 1998-2002 ('Towards better results-II' (Naar betere resultaten II)).[2]
This agreement was directly related to a restructuring of the governmental policies for disadvantaged groups that started in August 1998 (the so-called GOA-policy). A decentralization of policies towards municipal authorities was thought to be more effective in improving educational achievements of disadvantaged pupils at the local level. This restructuring provided Amsterdam with funds that were used in the second agreement. The new agreement extended the previous four-year plan with school specific performance targets and incentives. A short term target was formulated as increasing scores on the nationwide CITO-test for pupils in grade 8 of primary education to 532.7 points in the years 1999-2001. The long term target was set as scoring on average 534.6 points, which was the national average of 1998. In addition, specific targets were formulated for individual schools depending on their performance in the previous years and the socioeconomic composition of the school population measured in seven categories. For schools that in the previous three years scored below the average of their socioeconomic category, the target was set at the national average for this category of schools. For schools that already scored at the national average, the target was set as increasing the average scores with 2.5 points. Schools that already scored 2.5 points higher than the national average at least had to consolidate this performance. For reaching these targets the project included the following activities for schools in primary education:

[2] Municipality of Amsterdam, 1998, Governance agreement 'Towards better results 1998-2002' (Bestuursovereenkomst "Naar betere resultaten 1998-2002").

- All schools would participate each year in the CITO test for grade 8, this test would be used as an indicator for measuring school quality and performance;
- All schools would start to use systems for following the performance of individual students during primary education and use this system for yearly monitoring of the performance at the individual, group or school level;
- Each school would formulate a plan for the period 1998-2002 about the measures that would be taken for reaching the goals about improvement of the test scores;
- Each school would follow the municipal procedures for the assignment of pupils to different tracks in secondary education.

The municipality allocated additional funds to schools to carry out these activities. The funds obtained from the governmental GOA-policies were used. Each year, Amsterdam obtained approximately 12 million guilders from the GOA-funds, which totaled 175 million guilders for the whole country. At the school level this translated into on average 60,000 guilders for each school in Amsterdam (approximately 30,000 $ in 2000). The GOA-funds were supplemented with municipal resources. The allocation of the funds was made conditional on the implementation of the agreed activities and the realizations of the goals in each year. The plan formulated explicitly the rules for the allocation of funds which included a description of situations in which schools would not receive additional resources for disadvantaged pupils. For instance, schools that would not reach the targets and had not implemented the activities as agreed in their school plan would not receive the additional funds in the next school year. If the activities were implemented in the next year schools would again get the right to obtain the additional funds.

In 2001, nearly all schools (194 out of 195) participated in the CITO-test. In 2004 the project 'Towards better result' became part of a broader policy program for education and youth (Lokaal Onderwijs en Jeugdplan (LOJP). The scores on the CITO-test remained the main focus point.

Accountability policies in the rest of the country

Before 2000, there was no official accountability policy for The Netherlands as a whole. Although the Inspectorate of Education inspected schools and composed school reports, those reports were not available to the public. In 1998, several Dutch newspapers demanded the release of information about school performance. The newspapers used this information to compose ranking lists of *secondary* schools mainly located in the G4 and few other large cities.[3] In the following years secondary schools in other Dutch areas were also included in these rankings. In addition, newspapers started to publish rankings of schools in primary education. The fact that these lists were getting extensive public attention increased pressure on the

[3] The G4 cities are Amsterdam, Rotterdam, The Hague and Utrecht.

Inspectorate to make more information about schools available to the public. In 2003, the Inspectorate introduced on its website the “quality card”, which organizes information about a school in a compact and easily understandable manner. Before the introduction of the quality card, parents would have to download reports and read through them in order to find out how well a school fares on the CITO test.

3 Empirical strategy

Our main approach for assessing the effect of the municipal accountability policies of Amsterdam is to estimate standard difference-in-differences (DD) models. The first difference is the change in performance of pupils in Amsterdam before and after the introduction of the accountability policies. If pupils in Amsterdam perform better after the introduction of these policies this might be an effect of these policies. However, the improvement in performance might also be the result of other factors that changed in Dutch education during these years. To control for these other factors we use a second difference, which is the change in performance in a control group. This DD-approach rests on the assumption that the before-after difference for the control group would have been the before-after difference for the treated group in the absence of the reform. With the DD approach the treatment effect (β) of the Amsterdam accountability policies can be found as:

$$\beta = (Y^{A}_{after} - Y^{A}_{before}) - (Y^{C}_{after} - Y^{C}_{before}) \tag{3.1}$$

with Y^{A}_{after} is the performance of pupils in Amsterdam after the implementation of the accountability policies , Y^{A}_{before} is the performance of pupils in Amsterdam before the implementation of the accountability policies, Y^{C}_{after} and Y^{C}_{before} is the performance of pupils in the control group after and before the implementation of the Amsterdam policies. We estimate the treatment effect with a regression model which has the following form:

$$Y_{ist} = \beta_0 + \beta_1 A_{ist} + \beta_2 T + \beta_3 A_{ist}.T + \beta_4 X_{ist} + f_s + f_t + \varepsilon_{ist}, \tag{3.2}$$

where Y_{ist} is the performance on test Y of pupil i in school s in year t, A is a dummy which has value 1 if the person is a pupil in Amsterdam and value 0 if the person is a pupil in the control group (C), T is a dummy which has value 1 if the pupil took the performance test after the implementation of the Amsterdam policies and value 0 if the pupil took the performance test before the implementation of the Amsterdam policies, X is a vector of control variables, f_s and f_t are fixed effects for school and year of the survey, ε_{ist} is a person specific error term, and β is a vector of parameters to be estimated. The parameter of primary interest is β_3 which is the difference-in-differences estimator.

The data used in the analysis come from the so-called PRIMA-project in Dutch primary education (see next section). This project consists of two samples of schools. The first sample is representative for the Netherlands, the second sample includes an oversampling of pupils with a lower socioeconomic background. In our main estimation models we use these two samples as control groups, which means that the variable A has three categories. As the school population in Amsterdam includes large proportion of low SES pupils the second sample might be the most appropriate control group. The advantage of using these two control groups is that they measure

the nationwide change in performance for regular pupils and for pupils with lower socioeconomic backgrounds. The disadvantage is that these two samples might not pick up changes that occurred in the Dutch large cities. Therefore, we will also compare the change in educational performance in Amsterdam with the change in performance in the other three large Dutch cities ((Rotterdam, The Hague and Utrecht). Unfortunately, the number of schools from these cities decreased strongly in the last surveys. We therefore only use these schools for our robustness analyses. We estimated the standard DD-models for both the published test scores (CITO-test) and for the unpublished test scores (tests from the PRIMA-project). In addition, we also use DD-models for the analysis of the strategic behavior of schools.

An important issue in the analysis is the timing of the policies in relation to the observation window of our data. The first plan 'Towards better results' started in August 1995 and the second plan started in August 1998 (see section 2). The first plan was the initial phase of the new policies, the second plan set clear targets and was much more explicit about activities. The observation window of our main data stretches from early 1995 to early 2005 and includes biannual test scores (see next section). This means that the scores on the test taken in early 1995 are unlikely to be affected by the new policies. The scores on the test taken in early 1997 come from the initial phase of the new policies. It seems that the accountability policies really took form in the second plan. Therefore, in the difference-in-differences models we define the survey years 1995 and 1997 as the years before the treatment (T=0) and the survey years 1999, 2001, 2003 and 2005 as the treatment year (T=1).

4 Data

The data we use in the analysis were available from the longitudinal PRIMA survey. This biannual survey data is used to analyze the educational strategies and performance of the primary education system in the Netherlands (Driessen, van Lange, Vierke, 2004; Driessen, van Lange, Oudenhoven, 1994).We used the first six waves of the PRIMA survey including data on pupils, parents, teachers and schools from the school years 1994-95, 1996-97, 1998-99, 2000-01, 2002-03 and 2004-05. The PRIMA project consists of a panel of approximately 60,000 pupils in 600 schools. The participation in the project is voluntary. The main sample, which includes approximately 420 schools, is called the reference sample. An additional sample includes 180 schools for the over-sampling of pupils with a lower socioeconomic background (the Low SES sample). After each wave of the project some schools drop out and some new schools are included. However, there are no significant differences between the schools that drop out and the schools that remain in the project (Roeleveld and Vierke, 2003). Within each school, pupils in grades 2, 4, 6 and 8 (average age: 6, 8, 10, 12 years) are tested in language and arithmetic. Additionally, information on the social background is collected, and teachers are asked about the behaviour of the child in school. The nationwide CITO-test is taken independently from the PRIMA-project. At the end of the school year schools were asked to report the score of their pupils and these scores were added to the PRIMA data.

Dependent variables

The main dependent variables in the analysis are the scores on the CITO-test and the scores on the PRIMA-tests in language and arithmetic. The CITO-test is the most important nationwide test administered by over 80% of primary schools. The CITO-test is not compulsory. Every year, usually in February, pupils in their final year of primary education (grade 8, age 12) take the so-called *Eindtoets Basisonderwijs* test (CITO-test). The standardized test covers four areas:

- Language: spelling, writing, reading, and vocabulary;
- Arithmetic: understanding of numbers, mental arithmetic, percentages, fractions, dealing with measures, weights, money, and time;
- Information processing: use of texts, and other information sources, reading and understanding of tables, graphs, and maps;
- World orientation (optional): applying knowledge in the fields of geography, history, biology, science, and form of government.

The complete test consists of over 200 multiple-choice questions. The tests are comparable across years. CITO distinguishes five subtests. Unfortunately, our data do not contain the scores for the subtests of 1995. Therefore, we only analyze scores on the total test. Testing takes place over a period of three days in February. The outcome of the test is important for both pupils and

schools. Pupils' scores are used to help assign pupils to different levels of secondary education. The average scores of schools' pupils are used to judge the quality of primary schools. Parents use this information when choosing a primary school for their children. Every year the test receives considerable media attention, with national newspapers and television reporting on the most recent results. The primary aim of the CITO-test is to predict student success in secondary education.

The PRIMA-tests for languages and arithmetic were also developed by the CITO group but taken as part of the PRIMA-project (Kamphuis, Mulder, Vierke, Overmaat, Koopman, 1998). The language test for children in second grade, which is equivalent to infant school, measures the understanding of words and concepts. The arithmetic test for these children focuses on the sorting of objects. These tests can be taken in class. The test for children in grades 4, 6 and 8 all come from a system for following pupil achievements in primary education developed by the CITO group. The aim of these tests is to observe to which extend students master various elements of the curriculum. The tests for the same grade levels are identical each year. This ensures that the comparison of achievement levels over time is possible. The scores are also comparable between grades. The scales of the raw scores for language and arithmetic have no clear meaning. We have therefore opted to transform these scores for each test and each grade into wave specific standardized scores, having mean zero and standard deviation one. It should be noted that the comparability over time is hampered by other differences between waves. In the first wave, tests were taken early in the school year. In the second wave, tests were taken halfway through the school year. In the first two waves, tests were administered by an external examiner, while in the third wave the class teacher administered the tests. Because these differences may affect our findings we control for the year of the survey in all regressions.

Explanatory variables

All schools in the PRIMA-project have a school specific ID but the location of the schools cannot be identified from this ID. We obtained additional information on the municipality of the schools for identifying schools in Amsterdam and in the other three large Dutch cities (Rotterdam, The Hague, Utrecht). At the individual level we control for gender, age (in survey year, measured in days) and the pupil's so-called weight factor (subsidy factor) assigned by the funding scheme for primary schools. The Dutch funding scheme for primary schools distinguishes several groups of disadvantaged pupils. The most important groups are Dutch pupils with lower educated parents and pupils with an ethnic minority background. Pupils not belonging to a disadvantaged group enter the funding scheme with a weight factor equal to unity. Dutch pupils of poorly educated parents have a weight equal to 1.25 and pupils from an ethnic minority have a weight factor of 1.9. Schools receive 25 % additional funding for pupils

with a weight of 1.25, and 90% additional funding for these pupils with a weight of 1.90. Hence, this weight factor indicates the socio-economic background of the pupils.[4]

Main estimation sample

Our main estimation sample consists of pupils in grade 8 of primary education. Table 4.1 shows sample statistics of the main variables for the reference sample (column (1)), the Low SES sample (column (2)) and for Amsterdam (column (3)).

Table 4.1 Sample statistics of main estimation sample of pupils in grade 8 in 1995-2005 (standard deviations in brackets)

	(1) Reference Sample	(2) Low SES Sample	(3) Amsterdam
CITO score	534.3 (10.0)	529.3 (10.4)	529.9 (10.6)
Participation (%)	63.7	59.5	61.6
PRIMA			
Math score	0.12 (0.99)	– 0.23 (0.98)	– 0.24 (0.97)
Language score	0.16 (0.99)	– 0.33 (0.98)	– 0.36 (0.94)
Girl (%)	49.7	50.8	52.3
Age at test (years)	11.9	12.0	12.0
Subsidy factor (%)			
1.0	59.7	19.7	21.1
1.25	25.2	28.8	8.8
1.9	10.7	45.2	63.5
missing	3.8	5.2	6.6
Observations (schools)	54,169 (993)	21,534 (421)	5,698 (77)

Note: Not shown are subsidy factors 1.4 and 1.7 which include small proportions.

The main estimation sample consists of 81,401 pupils in 1,491 schools. The number of schools is more than 600 because after each wave schools drop out and new schools are included. The average scores on the CITO-test and on both PRIMA-tests are quite similar for pupils in Amsterdam and pupils in the Low SES sample, and clearly below the scores of pupils in the reference sample. The statistics for the subsidy factor show that there are major differences in the socio-economic background of pupils in Amsterdam and pupils in the Low SES sample compared to pupils in the reference sample, indicated by the subsidy factor. Although the Low SES sample is more comparable to the Amsterdam sample the former includes higher proportions of Dutch pupils with low educated parents (1.25) and smaller proportion of pupils from ethnic minorities (1.9). The sample of Amsterdam has been constructed from the two other samples. Three out of four pupils in the total sample of Amsterdam are drawn from the Low

[4] The PRIMA-project includes additional variables about the socioeconomic background of the pupils. However, these variables are not consistently measured over time. In addition, the school fixed effects in our models control for differences in the socioeconomic composition of the school population.

SES sample and one out of four pupils from the reference sample. Table A.1 in the appendix shows the averages of the covariates for each survey year.

In addition to the main estimation sample of pupils in grade 8, we also analyze data of pupils in grade 2, 4 and 6. These pupils did not participate in the CITO-test but did participate in the PRIMA-tests on math and language. Moreover, we use data from the so-called LEO-project, which is the predecessor of the PRIMA-project. The LEO-project, which collected data in 1988 and 1990, does not include the CITO-test but includes scores on tests for math and language similar to the tests used in PRIMA. Unfortunately, there is no information available on the location of the schools. However, we can identify schools in LEO that also participated in PRIMA. For both 1988 and 1990 we can identify 20 schools in Amsterdam. We use the data from the LEO-project to investigate the long term trend (Table 4.3).

Trends in unadjusted scores of the population and the sample

A comparison of the unadjusted scores on the CITO test in our samples with the scores from population statistics is shown in Table 4.2. The table shows the means for the population, the reference sample and the Low SES sample.

Table 4.2 CITO-scores in the population and in the PRIMA-samples (means)

CITO	1995	1997	1999	2001	2003	2005
The Netherlands						
Population	534.4*	534.5	534.2	534.9	534.7	534.5
Reference sample	534.9	534.5	534.1	534.4	534.7	533.5
Low SES sample	528.9	528.9	528.8	530.0	529.9	529.5
Amsterdam						
Population	529.4*	529.9	531.7	533.3	533.1	536.6**
Sample	525.7	528.2	531.0	530.9	531.2	531.1

* Measured in 1996; 1995 is not available; ** Without pupils with a low school advice;

The first rows in table 4.2 show that the average CITO scores in the Netherlands are quite constant, both in the population and in the two samples. The average scores in the Low SES sample are lower due to the oversampling of disadvantaged pupils. The CITO-scores of pupils in Amsterdam strongly increase according to the population statistics. Between 1997 and 2003 we observe an increase of 3.2 points. The scores for 2005 probably are inflated by the exclusion of weak pupils. The average scores in our ‘Amsterdam’ sample are lower because 75 % of the Amsterdam sample is drawn from the Low SES sample. Since 1997 we observe an increase of approximately 3.0 CITO-points. The unadjusted scores also suggest an increase of the CITO-scores between 1995 and 1997.

Trends in adjusted test scores in Amsterdam and in the rest of the country

For a first impression of the effect of the Amsterdam policies we compare the trend in test scores in Amsterdam with the trend in the two samples. We investigated whether the trend in the three test scores in Amsterdam diverged from the trend in the other two samples by estimating difference-in-differences models that include interaction of the year of survey and a dummy for Amsterdam and the full set of controls (age, age squared, subsidy factor, school fixed effects):

$$Y_{ist} = \beta_0 + \beta_1 A_{ist} + \beta_2 .Year + \beta_3 A_{ist} .Year + \beta_4 X_{ist} + f_s + \varepsilon_{ist}, \qquad (4.1)$$

Table 4.3 shows the estimates for the interaction variables in models for the CITO-test and the tests from the PRIMA-project. For the latter tests we also include data from the LEO-project which extends the observation window to 1988. Unfortunately, the LEO-project does not include scores on the CITO-tests.

Table 4.3 **Estimates of the changes in average scores in Amsterdam in grade 8 compared to the countrywide change by year of survey (1997=0)**

	1988	1990	1995	1997	1999	2001	2003	2005
CITO								
Amsterdam	n.a.	n.a.	0.003	0.0	0.543	0.461	0.461	0.492
			(0.112)	0.0	(0.092)***	(0.070)***	(0.084)***	(0.117)***
Low SES	n.a.	n.a.	0.070	0.0	0.133	0.139	0.067	0.090
			(0.057)	0.0	(0.051)***	(0.041)***	(0.054)	(0.060)
Observations								50840
Math								
Amsterdam	0.009	-0.159	-0.127	0.0	0.188	0.298	0.196	0.319
	(0.099)	(0.102)	(0.083)	0.0	(0.088)**	(0.072)***	(0.090)**	(0.096)***
Low SES	– 0.172	– 0.125	– 0.038	0.0	– 0.041	0.002	– 0.095	– 0.023
	(0.049)***	(0.047)***	(0.041)	0.0	(0.042)	(0.037)	(0.051)*	(0.061)
Observations								97274
Language								
Amsterdam	0.069	-0.129	-0.005	0.0	0.118	0.290	0.257	0.295
	(0.088)	(0.084)	(0.060)	0.0	(0.058)**	(0.048)***	(0.065)***	(0.064)***
Low SES	– 0.149	– 0.093	– 0.048	0.0	– 0.034	0.014	– 0.041	– 0.051
	(0.037)***	(0.034)***	(0.029)*	0.0	(0.030)	(0.025)	(0.036)	(0.037)
Observations								99801

Note: DD-estimates of the trend in test scores from regression models controlling for gender, subsidy factor, age and age squared, school fixed effects; standard errors adjusted for clustering at school year level in brackets.

The estimates in table 4.3 show that test scores in Amsterdam did not diverge from the trend in the Netherlands until 1997. However, from 1999 onwards we observe a clear improvement of the scores on all three tests compared to the general trend. This is the period of the second agreement between the schools and the municipality (Towards better results-II). For the Low SES sample we observe no improvement on the CITO-test and some improvement between the LEO-project (1988 and 1990) and the PRIMA-project that started in 1995.

In the next sections, we will only use the data from the PRIMA-project because of the consistency in the measurement of the three tests and the missing information on the location of schools in the LEO-project.

5 The effect of the urban policies on the published test

As a first step in the assessment of the increase in test scores in Amsterdam we analyze changes on the published test scores from the nationwide CITO-test taken by pupils in grade 8. We estimate difference-in-differences models of the effect of the Amsterdam policies on the scores of the CITO-test using different specifications for equation (2). The first specification in Table 5.1 (column (1)) does not include student characteristics or school fixed effects. As such, this specification can be seen as the closest proxy for the population statistic. This estimate is informative about possible sampling bias. The next two specifications (column (2) and (3)) show the robustness of the estimates for including different sets of controls. Hence, these estimates control for changes in the composition of the sample of students that took the test. The first three columns of Table 5.1 show the effects on the standardized CITO-test, the last column shows the effect on the CITO-score measured in points (500-550). The top panel shows the estimates for the whole sample of schools that participated in the CITO test, the bottom panel shows the estimates for the reference sample. Standard errors are corrected for clustering at the school year level.

Table 5.1 **Table 5.1 Difference-in-differences estimates of the effect of the Amsterdam policy on the CITO - test**

	Standardized CITO			CITO
	(1)	(2)	(3)	(4)
Amsterdam	0.454	0.509	0.490	5.189
	(0.090)***	(0.091)***	(0.070)***	(0.744)***
Low SES sample	0.124	0.137	0.089	1.003
	(0.045)***	(0.042)***	(0.039)**	(0.405)**
N	50840	50840	50840	50840
Controls				
Year dummies	Yes	Yes	Yes	Yes
Student characteristics	No	Yes	Yes	Yes
School fixed effects	No	No	Yes	Yes

Note: Column (1) controls for the cohort year, column (2) also controls for gender, age, age squared and subsidy factor, column (3) and (4) also include a school fixed effect.

The difference-in-differences estimates in column (1) show that test scores in Amsterdam increased with approximately 0.5 standard deviation after the introduction of the new urban policy. This estimate is very similar to the population statistic suggesting that sampling bias is not a serious concern. Including additional controls, such as the subsidy factor, slightly increases the estimates. The estimated increase corresponds to approximately 5-6 CITO-points (column (4) and (5)). The estimates in the bottom panel, using the reference sample, show a similar pattern. Test scores of pupils in Amsterdam increased with 0.4 to 0.5 standard deviation more than test scores of pupils in the rest of country. This suggests that the estimates are robust for changes in the composition of the sample of test takers. The second row in table 5.1 shows

that the performance of low SES students in the Netherlands improved with approximately 0.1 standard deviation. Hence, compared to this groups of students the performance in Amsterdam improved with 0.4 standard deviation.

The heterogeneity of the effects of the urban policy

Previous studies showed that the effects of accountability policies might differ between types of students and schools. For instance, Jacob (2005) finds that the accountability policy in Chicago especially affected marginal students and schools. This is consistent with the design of the policy in Chicago that imposed greater incentives on low-performing schools and students. The Amsterdam policy did not include clear differential incentives. However, schools might have chosen to focus their efforts on specific groups of students, for instance ethnic minorities or low performing students. We examined the heterogeneity of the effects of the urban policy by comparing the effects for different socioeconomic groups and by estimating quantile regressions (Table 5.2).[5] The top panel of table 5.2 shows the estimates of the main model (column (3) of Table 5.1) in which the urban policy is interacted with the subsidy factor based on the socioeconomic background of the pupils. The bottom panel shows the estimation results of the main model for various quantiles of the test score distribution.

Table 5.2 Difference-in-differences estimates of the effect of the Amsterdam policy on the standardized CITO –test for socioeconomic groups and quantiles of the test score distribution

	(1)	(2)	(3)	(4)	(5)
Quantile	10 %	25 %	50 %	75 %	90 %
Amsterdam	0.477	0.549	0.559	0.614	0.418
	(0.067)***	(0.058)***	(0.051)***	(0.047)***	(0.041)***
Low SES sample	0.138	0.166	0.131	0.122	0.059
	(0.038)***	(0.033)***	(0.029)***	(0.026)***	(0.023)**
N	50840	50840	50840	50840	50840
		Regular Dutch	Low educated	Ethnic minority	
Amsterdam*socio-economic group	0.510	0.160	0.500		
		(0.116)***	(0.119)	(0.080)***	
Low SES * socioeconomic group	-0.033	-0.040	0.212		
		(0.059)	(0.045)	(0.048)***	
N				50840	

Note: The top panel shows estimates of quantile regressions for 5 quantiles. The bottom panel shows estimates of an interaction of subsidy factor with Amsterdam or Low SES sample from the main model in column (3) of table 5.1.

[5] Unfortunately, we cannot investigate the heterogeneity of the effect of policy between schools in Amsterdam because of the limited number of schools that participated in more than one survey of the PRIMA-project.

The top panel shows that the Amsterdam policies affected the performance of pupils in all quantiles of the test score distribution. At the tales of the distribution the size of the estimates is slightly lower. For the low SES sample the estimates are quite similar for all quantiles, but smaller in the highest quantile. The bottom panel shows that the improvement in test scores in Amsterdam is similar for the two largest groups in Amsterdam: the regular Dutch pupils and for ethnic minorities. For pupils with lower educated Dutch parents the estimate is statistically not significant. For the Low SES sample we find only an improvement of test scores for students from ethnic minorities. This suggest that the total improvement in Amsterdam compared to the other two samples is primary driven by the regular Dutch pupils and to a lesser extent by students from ethnic minorities.

6 Teaching to the test?

One of the main lessons from the recent economic literature on school accountability is that increases in test scores might be the result of strategic reactions of schools (Jacob, 2005). For instance, schools might practice a lot with tests from previous years. This might increase test-specific skills but might not improve general skills. This is often labeled as 'teaching-to-the-test'. In this section we investigate this key question for the assessment of the increase in test scores in Amsterdam in two ways. First, we investigate the change in performance of pupils in Amsterdam on a second independent test (the PRIMA-test). If the accountability policies improved the general skills of pupils we expect the increase in scores on the CITO-test to be reflected in an increase of scores on the PRIMA-test. Second, we analyse the change in educational performance of pupils in earlier grades. The accountability policies only set targets for the performance of pupils in grade 8.

6.1 The effect of the urban policy on the unpublished tests

We investigate the change in learning abilities by estimating difference-in-differences models on scores on specific tests from the PRIMA-project. The key difference is that these tests have not been published and do not play a role in the accountability policies. Table 6.1 shows the estimates of the effect of the Amsterdam accountability polices on the tests from the PRIMA project taken in grade 8. The first three columns show the estimates for the math test using different sets of controls, the last three columns show the estimates for the language test. The top panel shows the estimates for the sample of schools that participated in the CITO test. Hence, this sample is comparable to the sample used in Table 5.1. The bottom panel shows the estimates for the total sample of the PRIMA-project.

When using the sample of 'CITO-participants' we find for both (unpublished) tests that the scores in Amsterdam increased with approximately 0.3 standard deviation compared to the change in the rest of the country. The sample of CITO-participants consists of schools that participated in the CITO-test and also includes scores of pupils that did not participate in the CITO-test. If we restrict the sample to pupils that participated in the CITO-test we find similar results. The results for the sample of PRIMA-participants are also similar. For the pupils in the Low SES sample we find no improvement of performance in the model that includes all controls.

Table 6.1 **Difference-in-differences estimates of the effect of the Amsterdam policy on the PRIMA tests in math and language in grade**

	PRIMA Math			PRIMA Language		
	(1)	(2)	(3)	(4)	(5)	(6)
CITO participants						
Amsterdam	0.286	0.315	0.373	0.319	0.366	0.327
	(0.084)***	(0.084)***	(0.073)***	(0.068)***	(0.058)***	(0.056)***
Low SES sample	0.022	0.033	-0.072	0.072	0.082	0.003
	(0.042)	(0.040)	(0.045)	(0.039)*	(0.031)***	(0.031)
N	50058	50058	50058	51858	51858	51858
PRIMA participants						
Amsterdam	0.267	0.278	0.308	0.296	0.311	0.239
	(0.073)***	(0.071)***	(0.067)***	(0.063)***	(0.047)***	(0.049)***
Low SES sample	0.074	0.089	– 0.010	0.085	0.098	0.021
	(0.037)**	(0.035)**	(0.032)	(0.031)***	(0.024)***	(0.024)
N	74726	74726	74726	77246	77246	77246
Controls						
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Student characteristics	No	Yes	Yes	No	Yes	Yes
School fixed effect	No	No	Yes	No	No	Yes

Note: Column (1) and (4) control for year dummies, column (2) and (5) also control for gender, age, age squared and subsidy factor, column (3) and (6) also include a school fixed effect.

General skills versus test-specific skills

The improvement of pupils in Amsterdam on the PRIMA-tests is in line with the findings on the change in performance on the CITO-test, and suggests that the improvement of Amsterdam pupils on the CITO-test reflects an improvement of general skills. We decomposed the increase of scores on the CITO-test in a test-specific and a general component by including the PRIMA-test scores as controls in the models of the previous section. The estimates of the regression on the standardized score on the CITO-test are shown in table 6.2. Including the PRIMA-test scores as controls reduces the sample due to missing values on these tests. Column (4) re-estimates the main model of table 5.1 (column (3)) using this smaller sample.

The estimates for the total sample suggest that approximately 60 percent of the total increase in performance on the CITO-test score is driven by an increase in general skills and 40 % by an increase in test-specific skills. The estimates for the reference sample suggest that more than 80 percent can be attributed to an increase in general skills. For the Low SES sample the inclusion of the PRIMA test scores does hardly change the estimates. This suggests that the total improvement for this sample is test-specific.

Table 6.2 **Difference-in-differences estimates of the effect of the Amsterdam policy on the CITO –test controlling for the scores on the PRIMA-tests**

	Standardized CITO			
	(1)	(2)	(3)	(4)
Amsterdam	0.193	0.210	0.194	0.527
	(0.044)***	(0.045)***	(0.053)***	(0.073)***
Low SES sample	0.099	0.104	0.116	0.077
	(0.027)***	(0.027)***	(0.028)***	(0.040)*
N	46364	46364	46364	46364
Controls				
PRIMA test scores	Yes	Yes	Yes	No
Year dummies	Yes	Yes	Yes	Yes
Student characteristics	No	Yes	Yes	Yes
School fixed effects	No	No	Yes	No

Note: Column (1) controls for the cohort year, type of sample and G3 city, column (2) also controls for gender, age, age squared and subsidy factor, column (3) and (4) also include a school fixed effect.

6.2 The performance of pupils in grade 2, 4 and 6

Previous research has found that accountability policies only improved test scores in the high stakes grade (Figlio and Rouse, 2006). This finding might indicate teaching to the test. The PRIMA project also measures the cognitive ability of pupils in grade 2, 4 and 6. This provides the opportunity to investigate whether the Amsterdam policies also had an effect on the performance in earlier grades. If the performance of pupils in these grades improved it seems not likely that this improvement is the result of special practicing for taking the CITO-test in grade 8. Table 6.3 shows the estimates from DD-models similar to column (2) and (3) of Table 6.1. The left panel shows the estimated effect on scores in math, the right panel shows the estimates for languages. We show the results for the pooled sample of grade 2, 4 and 6 while controlling for grade, and the results for the separate grades.

The results in Table 6.3 indicate that the accountability policies in Amsterdam not only increased test scores of pupils in grade 8 but also in earlier grades. For the pooled sample the test scores of pupils in Amsterdam increased 0.1 to 0.2 standard deviation more than in the rest of the country. For the Low SES sample we observe a decrease of test scores of approximately 0.1 standard deviation in math and no improvement in language. For the separate grades all point estimates of the effect of the Amsterdam polices are positive but not all estimates are statistically significant. For the low SES sample there seems to be a deterioration of test scores. Although the size of the improvement in performance is smaller than the improvement in grade 8 these estimates suggest that the Amsterdam policies also increased the general skills of pupils in earlier grades.

Table 6.3 Estimates of the effect of the Amsterdam policies on pupils in grade 2, 4 and 6

	Math		Language	
	(1)	(2)	(3)	(4)
Grade 2-6				
Amsterdam	0.181	0.133	0.161	0.119
	(0.044)***	(0.054)**	(0.052)***	(0.049)**
Low SES	– 0.004	– 0.097	0.055	– 0.017
	(0.023)	(0.022)***	(0.022)**	(0.019)
N	248893	248893	250282	250282
Grade 2				
Amsterdam	0.285	0.228	0.171	0.129
	(0.071)***	(0.101)**	(0.080)**	(0.095)
Low SES	0.066	– 0.006	0.040	– 0.039
	(0.035)*	(0.035)	(0.033)	(0.034)
N	85765	85765	85217	85217
Grade 4				
Amsterdam	0.089	0.101	0.106	0.056
	(0.066)	(0.077)	(0.068)	(0.067)
Low SES	– 0.059	– 0.160	0.044	– 0.027
	(0.031)*	(0.033)***	(0.031)	(0.031)
N	84750	84750	84865	84865
Grade 6				
Amsterdam	0.175	0.047	0.222	0.144
	(0.060)***	(0.063)	(0.064)***	(0.050)***
Low SES	– 0.024	– 0.117	0.079	0.006
	(0.032)	(0.029)***	(0.028)***	(0.026)
N	78378	78378	80200	80200
Controls				
Year dummies	Yes	Yes	Yes	Yes
Student characteristics	Yes	Yes	Yes	Yes
School fixed effect	No	Yes	No	Yes

Note: Column (1) and (3) control for cohort year, gender, age, age squared and subsidy factor, column (2) and (4) also includes school fixed effects.

7 Shaping the testing pool

The economic literature provides evidence that accountability policies might induce schools to shape the testing pool by excluding pupils from the test (see introduction). In this section, we investigate various channels for shaping the testing pool: direct exclusion, assignment to special education, retention and exploiting exceptions of the testing rules (giving low advices for secondary education).

7.1 Excluding weak pupils

To investigate the direct exclusion of pupils from the CITO-test we start by analysing the changes in the number of pupils for which we do not observe a score on the CITO-test in our sample. It should be noted that a missing value on the CITO-test may have many reasons, such as non-reporting, illness of the pupils at the time of the test or strategic behaviour of schools. Table 7.1 shows the proportion of missing values on the CITO-test by location for each survey year for the sample of schools and classes that reported at least one score on the CITO test.[6]

Table 7.1 Missing values on the CITO-test (%) by sample and year of survey in schools that participated in the CITO-test

	1995	1997	1999	2001	2003	2005
Amsterdam (%)	5.4	11.1	13.0	8.4	9.6	5.9
N	591	440	414	680	876	834
Low SES sample (%)	7.4	5.5	6.0	8.3	3.5	9.4
N	2015	1866	2339	2799	2458	2266
Reference sample (%)	3.6	5.1	4.8	5.5	5.1	8.6
N	4984	5675	5915	6442	6698	6839

If the accountability policies induced strategic behaviour we might observe relatively more missing values in schools in Amsterdam than in schools in the other two samples after the implementation of the policies. However, we do not observe such a pattern. In Amsterdam the proportion of pupils with a missing value on the CITO-test increased until 1999 but decreased afterwards. For the other samples we also do not observe a clear pattern.

In Table 7.2 we take a closer look at these changes. The first column shows the difference-in-differences estimate of the effect of the Amsterdam policies on the probability of not taking the CITO-test. The estimate shows that after the implementation of the policies the probability of not taking the test decreased in Amsterdam compared to the rest of the country with 0.4 percentage points, which is statistically insignificant. For the Low SES sample the decrease is 1.2 percentage points. In addition, we checked for changes in the composition of test-takers by

[6] For some schools that participated in the CITO-test the scores of all pupils within a class are missing. It seems likely that the teachers failed to send the CITO-scores of their classes to the PRIMA-project. We consider these missing values as a non-reporting issue and exclude them from the analysis.

exploiting the fact that we have scores on two independently taken tests. We compared the scores on the PRIMA-test of the pupils that took the CITO-test and the pupils that did not take the CITO-test. For 92 (89) percent of pupils with a missing CITO-score we observe a score on the PRIMA language (math) test. Column (2) and (3) show the difference-in-differences estimates of the effect of having a missing CITO-score on the scores on the PRIMA-test. The DD-estimates for pupils in Amsterdam are positive and statistically not significant. Hence, excluded pupils in Amsterdam after the introduction of the accountability polices did not score lower but slightly higher on the PRIMA-tests. In sum, the estimates in Table 7.2 do not provide evidence that the increase in test scores in Amsterdam can be explained by the direct exclusion of (weak) pupils from the CITO-test.

Table 7.2 Difference-in-differences estimates of the direct exclusion effect

	Missing CITO	P-math	P-language
	(1)	(2)	(3)
Amsterdam	– 0.004	0.047	0.014
	(0.022)	(0.167)	(0.144)
Low SES sample	– 0.012	– 0.001	– 0.071
	(0.015)	(0.125)	(0.113)
N	54135	49834	51628

Note: Column (1) shows coefficients of a DD-model on having a missing CITO-score, column (2) and (3) show estimates of DD-models of a missing CITO-score on the PRIMA-tests; all models control for year dummies, student characteristics and include school fixed effects.

7.2 Other strategic behavior

Assignment to special education

Previous studies show that accountability policies might induce schools to assign more students to special education (Jacob, 2005) or to classify more students as disabled (Figlio and Getzler, 2006). We investigate whether this strategic behavior also occurred in Amsterdam by looking at the change in the number of pupils in special education. Table 7.3 shows the proportion of pupils of all grades in regular and special education by sample in the period 1995-2005. Special education consists of two types: special primary and special education.

Table 7.3 Proportion of pupils in regular and special primary education 1995-2005

CITO		1995	1997	1999	2001	2003	2005
Netherlands	Regular	94.8	94.9	95.0	95.0	94.8	94.8
	Special	5.2	5.1	5.0	5.0	5.2	5.2
Amsterdam	Regular	92.9	93.2	93.3	93.4	93.3	93.3
	Special	7.1	6.8	6.7	6.6	6.7	6.7

The first rows of Table 7.3 shows that the proportion of pupils in some type of special education in the Netherlands is approximately 5.2 % and quite stable over time. For Amsterdam we observe a decrease of the proportion of pupils in special education from 7.1 % in 1995 to 6.7 % in 2005. Considering this downward trend in Amsterdam it seems not likely that schools used the assignment of pupils to special education to increase scores on the CITO test.

Retention

An increase of retention might also be a channel for shaping the test pool. An additional year in primary education might increase test scores of weak pupils. We investigated this channel by comparing the age of pupils in grade 8 before and after the introduction of the accountability policies in Amsterdam and in the other samples. Difference-in-difference estimates using the same specifications as in the models of the previous sections yield negative and statistically insignificant estimates for Amsterdam. Hence, it seems not likely that Amsterdam used this channel for improving test scores.

Not reporting test scores of pupils with a low advice for secondary education

In 2004 the municipality of Amsterdam decided that pupils with a low advice for secondary education no longer had to take the CITO-test[7]. Rotterdam took the same decision. This decision was based on the argument that the results on the CITO test did not have a predictive value for the school career of these pupils. The Dutch media suggested that the exclusion of these pupils raised test scores in Amsterdam. As schools themselves give school advices to their pupils there is scope for strategic behaviour: by issuing more low school advices schools can raise their own average test scores. In addition, schools do not have to report the results of these pupils on the CITO-test to the Inspectorate of Education. This provides schools with the opportunity to exclude the scores of certain students after they took the test.

However, it seems not likely that this second channel affects our previous results. First, the previous estimates of the performance on the CITO-test are based on the total sample of students including pupils who received a low advice for secondary education. In addition, the analysis in the previous section showed that the exclusion of pupils cannot explain the increase in test scores in Amsterdam. Second, we also found an improvement of the scores on the other tests for pupils in Amsterdam. It is not clear why schools would exclude pupils from an unpublished test. Third, the municipality of Amsterdam started to exclude pupils with low school advices in 2004. However, the increase in performance of pupils in Amsterdam can already be observed in the years before 2004. We have no evidence that Amsterdam already excluded pupils with a low school advice before 2004. To check this we compared the number of pupils in the municipal reports of Amsterdam with the number of pupils that took the test according to population data for the period 1999-2003. This comparison showed that the

[7] Specifically, pupils with advices for the so-called Practice Education (Praktijk onderwijs, PRO) and the School Career Supporting Education (Leerweg Ondersteunend Onderwijs, LWOO).

Amsterdam reports are based on all pupils that took the test which means that they did not exclude the (low) scores of students after they took the test.
In sum, we find no evidence that Amsterdam schools increased test scores by exploiting these channels of strategic behavior.

8 Robustness analysis

To further test the robustness of the results from the previous results we performed several sensitivity tests. First, inspection of the covariates in the appendix (see Table A.1) shows that a relatively large proportion of pupils in the Amsterdam sample of 1995 has a missing value on the subsidy factor. In Table 5.1 these pupils are included in the estimation with a dummy variable for missing on the subsidy factor. We excluded these pupils to test the sensitivity of the results. The estimates slightly decrease towards 0.448 in the model of column (3) after the exclusion of these pupils.

Second, we checked whether the change in the participation rules for the CITO-test by the municipality of Amsterdam in 2004 might affect our findings by excluding all observations from 2005. We find that the estimates do not change after excluding all observations from 2005.

Third, at the start of the first plan 'Towards better results' approximately 30 Montessori schools refused to participate in the CITO-test. They started taking the test in the school year 1997-1998 (the CITO-test is taken in February). Hence, they started participating at the cut-off between the first and second plan. We investigated the sensitivity of the results by excluding all schools (10 schools) in Amsterdam that started participating in the next PRIMA-project held in 1999. From our data we cannot observe whether schools are Montessori schools. After the exclusion of the schools in Amsterdam that entered the PRIMA-project in 1999 we find a slightly higher estimate of the effect on the CITO-score (0.492 (0.070). Hence, the results are robust for the participation of the Montessori schools.

Fourth, we investigated a potential 'large city effect'. In the previous analysis we used two nationwide samples as control group. A concern with these control groups is that they might not pick up changes in performance in Dutch large cities. To investigate this large city effect we compared the change in educational performance in Amsterdam with the change in performance in the other three large Dutch cities (Rotterdam, The Hague and Utrecht). Unfortunately, the number of schools from these cities in the PRIMA-sample decreased strongly between 1995 and 2005.[8] Hence, the comparison between Amsterdam and the other three large cities might suffer from sampling bias. Table 8.1 shows DD-estimates of the Amsterdam policy for models in which the other three large cities are taken as the control group. The left panel shows the estimates for the three tests taken in grade 8, the right panel shows the estimates for the PRIMA test taken in grade 2, 4 and 6.

[8] The total number of schools in the sample located in these three cities decreased steadily from 83 in 1995 to 19 in 2005.

Table 8.1 **Difference-in-differences estimates of the effect of the Amsterdam policy using the other three large cities as control group**

	Grade 8			Grade 2-6	
	CITO	Math	Language	Math	Language
	(1)	(2)	(3)	(4)	(5)
Amsterdam	0.235	0.244	0.264	0.193	0.165
	(0.088)***	(0.073)***	(0.063)***	(0.052)***	(0.055)***
N	50840	74726	77246	248893	250282

Note: All models include the same controls as used in the full model in column (3) of table 5.1.

The estimates show that for pupils in grade 8 the test scores in Amsterdam improved 0.2 to 0.3 standard deviation more than in the other three large cities. The size of the improvement on the CITO-test is similar to the size of the improvement on the PRIMA-tests. This suggests that the improvement of general skills is approximately 0.2 to 0.3 standard deviation which is similar to the findings from the comparison with the two nationwide samples. For pupils in grade 2 to 6 the improvement in Amsterdam is 0.2 standard deviation larger than in the other cities.
Although the estimates of the effect of the Amsterdam policy on the CITO-test are smaller than the estimates found in the previous section the findings on the improvement in the PRIMA-tests corroborate a substantial improvement of the general skills of pupils in Amsterdam in all grades of primary education.

9 Conclusion and discussion

In 1995 the municipality of Amsterdam introduced accountability policies for schools in primary education. Populations statistics show a remarkable increase of test scores after the introduction of the new urban policies. This paper assesses this increase in test scores by analyzing data of a large sample of schools. Our main finding is that after the introduction of the accountability policies test scores in Amsterdam increased more than in two nationwide samples. Scores of pupils in Amsterdam increased with 0.4 to 0.5 standard deviation more than in the reference sample and with 0.4 standard deviation more than in the Low SES sample. This increase confirms the findings from the population statistics.

We investigated whether the increase in the published test scores was based on an improvement of general skills or an improvement in test-specific skills by analyzing the scores on two other independently taken tests. The estimates show that pupils in Amsterdam also increased their performance on the other two tests but the improvement on these test is smaller than the improvement on the CITO-test. Both in math and in languages average test scores in Amsterdam improved with 0.3 standard deviation more than in the other two nationwide samples. A decomposition of the increase in scores on the CITO-test suggests that 60 percent of the increase is driven by an increase of general skills and 40 percent by an increase in test-specific skills. In addition, we investigated the test scores of pupils in earlier grades. The accountability policies only set targets for the performance of pupils in grade 8. We find that the performance of pupils in Amsterdam in earlier grades also increased more than the performance of pupils in other locations but the improvement in earlier grades is smaller than the improvement in grade 8.

We investigated various channels for strategic behavior of schools: the direct exclusion of pupils, assigning more pupils to special education, retention or exploiting exception from the test taking rules (giving more low school advices). However, we do not find evidence that the use of these channels can explain the increase in test scores in Amsterdam. A robustness check focused comparing the change in performance in Amsterdam with the change in performance in the other three large Dutch cities confirms the main findings that the Amsterdam policy improved general skills of pupils.

Previous research showed that accountability policies can improve test scores but schools will also try to improve their results through strategic behaviour. Against this background the results of the accountability policies in Amsterdam seem relatively successful.
This raises the question which components of the Amsterdam policy are important for these positive results. To get insight into this question we discussed our empirical findings with two directors of primary schools in Amsterdam. Both directors recognised the improvement in test scores after the introduction of the policies. Their main explanation for the improvement was that the accountability policies created a culture within schools that was more oriented on measuring and monitoring of performance. The directors of schools used this information of

pupils to support and monitor the performance of teachers. Teachers were more careful and put more effort in following the individual progress of pupils and started acting when the progress was too small. Visser (2003) in his examination of the school choice procedure also notes that more than half of the schools in Amsterdam reported in 1999 that they spent more time on basic skills in grade 7 and 8. In addition, 40 percent of the schools put more effort in differentiating between pupils instead of teaching at the class level. In our view two other components of the Amsterdam policy might also have been important. Firstly, the new policy set clear school specific short term and long term performance targets and not only focussed on low performing schools or pupils. Secondly, the policy was directly related with changes in admission rules for tracks in secondary education. The formulation of test score band widths for specific tracks in secondary education might have created additional incentives for individual students and their parents to put more effort in their education.

We conclude that our overall assessment of the accountability policies in Amsterdam is positive. Although part of the gains in test scores might be test-specific the policies seem to have led to a substantial improvement of general skills of pupils in Amsterdam.

References

Dee, T.S. and Jacob, B.A., 2009, The Impact of No Child Left Behind on Student Achievement. NBER Working Paper #15531.

Figlio, D. and with Lawrence Getzler, 2006, Accountabiliy, ability and disability: Gaming the system? In *Advances in Microeconomics, Vol. 14: Improving School Accountability - Check-ups or Choice?*, ed. T. Gronberg and D. Jansen, 35-49, Amsterdam: Elsevier.

Figlio, D., 2006, Testing. Crime and Punishment,, *Journal of Public Economics* 90(4-5): 837-51.

Figlio, D. and C. Rouse, 2006, Do accountability and voucher threats improve low-performing schools?, *Journal of Public Economics* 90(1-2): 239-55.

Figlio, D. and J. Winicki, 2005, Food for thought? The effects of school accountability plans on school nutrition, *Journal of Public Economics* 89(2-3): 381-94.

Jacob, B., 2005, Accountability, Incentives and Behavior: Evidence from School Reform in Chicago, *Journal of Public Economics*, 89(5-6): 761-796.

Jacob,B.A. and S.D. Levitt, 2003, Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating, *Quarterly Journal of Economics*, 118(3), 843-877.

Kamphuis, F., L. Mulder, H. Vierke, M. Overmaat, P. Koopman, 1998, *De relatie tussen PRIMA-toetsen en toetsen uit het CITO-leerling volgsysteem*, Nijmegen: ITS.

Municipality of Amsterdam, 1994, Towards better results (Naar betere resultaten, Een plan voor het onderwijs in Amsterdam 1994-1998).

Municipality of Amsterdam, 1998, Governance agreement ‘Towards better results 1998-2002’ (Bestuursovereenkomst “Naar betere resultaten 1998-2002”).

Roeleveld, J. and R. Portengen, 1998, *Uitval en instroom bij het Prima-cohortonderzoek*, Amsterdam/Nijmegen: SCO-Kohnstamm Instituut / ITS.

Roeleveld, J. and H. Vierke, 2003, *Uitval en Instroom bij de derde meting van het PRIMA-cohortonderzoek*, Amsterdam/Nijmegen: SCO-Kohnstamm Instituut / ITS.

Statistics Amsterdam, 1998, Key figures of Amsterdam (Amsterdam in cijfers), Yearly report 1997, Statistics Amsterdam (O&S).

Visser, M., 2003, School Choice Procedure, Quality and Quality Assurance (Schoolkeuzeprocedure, kwaliteit en kwaliteitszorg), Thesis, University of Groningen, the Netherlands.

Appendix 1

Table A.1 Means of test scores and socio-economic background by sample and year of survey

	1995	1997	1999	2001	2003	2005
Prima Math						
Reference	0.19	0.12	0.13	0.09	0.12	0.06
Amsterdam	– 0.44	– 0.33	– 0.23	– 0.19	– 0.26	– 0.07
Low SES	– 0.25	– 0.25	– 0.28	– 0.20	– 0.26	– 0.17
Prima Language						
Rest	0.24	0.17	0.19	0.15	0.14	0.11
Amsterdam	– 0.50	– 0.57	– 0.42	– 0.30	– 0.33	– 0.20
G3	– 0.34	– 0.35	– 0.39	– 0.32	– 0.29	– 0.28
Controls						
Age						
Rest	11.7	11.9	11.9	12.0	12.0	12.0
Amsterdam	11.6	12.1	12.1	12.1	12.1	12.0
G3	11.8	12.0	12.1	12.1	12.1	12.1
Gender (Girls)						
Rest	0.50	0.50	0.50	0.50	0.50	0.50
Amsterdam	0.55	0.48	0.52	0.50	0.54	0.52
G3	0.52	0.51	0.51	0.50	0.50	0.50
Socio-economic subsidy factor						
Reference						
1.0	48.9	52.2	54.8	63.9	67.7	68.4
1.25	35.5	33.2	28.4	22.4	18.8	15.5
1.9	8.6	9.9	10.3	10.7	11.3	13.1
Missing	6.4	4.1	5.6	2.4	2.0	2.7
Amsterdam						
1.0	11.7	16.7	17.4	25.5	23.1	27.3
1.25	12.3	11.9	11.0	7.9	6.8	5.8
1.9	52.7	67.6	64.7	64.1	69.2	62.9
Missing	23.1	3.7	6.9	2.5	0.9	4.0
Low SES						
1.0	12.1	14.3	14.1	22.9	30.5	28.6
1.25	34.6	32.9	29.3	25.3	24.3	24.0
1.9	42.7	44.5	50.5	47.8	41.9	44.3
Missing	9.6	7.4	5.2	3.0	2.0	2.5