CPB Discussion Paper

No 83

June 2007

Do School Inspections Improve Primary School Performance?

Rob Luginbuhl, Dinand Webbink, and Inge de Wolf[a]

[a] Dutch Inspectorate of Education

The responsibility for the contents of this CPB Discussion Paper remains with the author(s)

CPB Netherlands Bureau for Economic Policy Analysis
Van Stolkweg 14
P.O. Box 80510
2508 GM The Hague, the Netherlands

Telephone +31 70 338 33 80
Telefax +31 70 338 33 50
Internet www.cpb.nl

ISBN 978-90-5833-325-4

Abstract in English

Inspectors from the Dutch Inspectorate of Education inspect primary schools, write inspection reports on each inspected school, and make recommendations as to how each school can improve. We test whether these inspections result in better school performance. Using a fixed-effects model, we find evidence that school inspections do lead to measurably better school performance. Our assessment of school performance is based on the Cito test scores of pupils in their final year of primary school. Therefore school improvement means increased Cito test scores. The results indicate that the Cito test scores improve by 2% to 3% of a standard deviation of the test score in the two years following an inspection. The arithmetic component shows the largest improvement. Our estimates are the result of an analysis of two types of school inspections performed between 1999 and 2002, where one type was more intensive than the other. In one fixed-effects model, we assume that the effect of the two types of school inspections was the same. We cannot, however, be sure that the estimates from this model are free from the problem of endogeneity bias. Therefore, we also obtain estimates for a less restrictive fixed-effects model. In this less restrictive model, we make use of the fact that a subset of the more intensive school inspections occurs at a representative selection of primary schools. Based on this smaller, essentially randomly drawn sample of schools, we can be confident that these estimates of the effect of school inspections are free from endogeneity bias. Due to the limited number of inspections at randomly selected schools, these estimates are not significantly different from zero. These estimates are, however, consistent with the effects found based on all inspections. The less restrictive model also allows for the effect of the more intensive inspections to differ from that for the less intensive ones. We find evidence that the more intensive inspections are responsible for larger increases in the Cito test scores than the less intensive ones.

Abstract in Dutch

De Inspectie van het Onderwijs bezoekt scholen in het primair onderwijs, schrijft op basis daarvan een inspectieverslag en doet aanbevelingen om de prestaties te verbeteren. Deze studie onderzoekt het effect van de schoolbezoeken van de Inspectie van het onderwijs aan scholen in het basisonderwijs. Daarvoor wordt de verandering in de Citoscores vergeleken van scholen die bezocht zijn door de Inspectie met de verandering bij scholen die niet zijn bezocht door de Inspectie. De analyse is gebaseerd op een bestand van alle Citoscores in Nederland over de jaren 1999-2003. In de analyse wordt onderscheid gemaakt tussen kortere en intensievere bezoeken aan scholen. In de eerste twee jaren na het inspectiebezoek stijgen de Citoscores met ongeveer 2% tot 3% van een standaarddeviatie. De stijging is groter voor de intensievere bezoeken en in rekenen. In de analyse is ook gekeken naar de verandering in een random steekproef die door de Inspectie is getrokken voor het maken van het jaarlijkse Onderwijsverslag. Het voordeel van

deze steekproef is dat deze aselect is getrokken, nadeel is dat het aantal bezoeken veel kleiner is. Deze analyse geeft kleinere positieve effecten van inspectiebezoeken, die evenwel consistent zijn met de eerdere bevindingen. Voor rekenen worden ook met deze aanpak positieve effecten gevonden die statistisch significant zijn.

Contents

Summary

A number of countries, mainly European ones, have governmental agencies to inspect schools. The Dutch Inspectorate of Education inspects primary schools to ensure that the schools are complying with Dutch educational laws. These inspections are, however, also intended to improve the quality of the education the schools provide.

Dutch school inspectors write a report on each inspected school following an inspection and make recommendations as to how each school can improve. In this paper we ask what effect, if any, these school inspections have on the test scores of Dutch primary school pupils.

Using a number of model specifications, we find evidence that school inspections do lead to measurably better school performance. Our assessment of school performance is based on the Cito test scores of pupils in their final year of primary school. Therefore school improvement means increased Cito test scores.

Our main finding is that school inspections lead to better performance of schools. In the first two years following an inspection test scores increase by 2% to 3% of a test score's standard deviation. The improvement in Dutch elementary schools is strongest in the area of arithmetic and persists over the four years following an inspection that our data covers. For the three other subject areas covered by the Cito test and for the test score total, the improvement is significant in the two years following an inspection. Thereafter the estimated effects are typically positive and of similar magnitude to those in the first two years, but not significant.

In this paper we have also been able to look into the effectiveness of two different types of inspections. The Dutch Inspectorate of Education carries out two types of inspections which differ in intensity. The less intensive version (RST) takes approximately one day and the more intensive version (IST) takes 2 to 3 days. In one of our model specifications, we assume that the effect of these two types of school inspections was the same. In an alternative specification, we allow the effect of a school inspection to differ depending on whether the inspection was a less intensive RST inspection, or a more intensive IST one. Our analysis also indicates that the more intensive inspections produce larger improvements in school performance than the less intensive ones.

Estimating the effect of school inspections on primary school performance is difficult, because inspectors may not randomly select which schools they inspect. School inspectors are likely to visit poorly performing schools more often. They may also inspect schools for reasons a researcher can not observe. This non-random selection can produce an endogeneity bias in the estimates of the impact of school inspections. As a result, an estimated effect could actually be due to correlation between unobserved heterogeneity in the quality of schools and the inspectors' decisions about which schools to inspect.

We use two approaches to overcome this bias. First, we estimate models that include fixed school effects using test scores over the period 1999 to 2003. These models control for

unobserved school factors that are time invariant. This ensures that correlation between the inspections and constant factors, such as the physical infrastructure of the school, will not bias the estimates. In this approach only changes in a school's quality over time can bias our estimates if they are correlated with (but not caused) by the school inspections.

Our second approach also tries to rule out this possible bias. We exploit the fact that the Dutch Inspectorate of Education inspects a random sample of schools in order to compile their annual report on the state of Dutch education. In doing so, the inspectorate performs what is essentially a controlled experiment in which the sample of schools that are inspected represents a random experimental group. The advantage of this sample is that we can be more confident about the estimates. However, the number of inspections at randomly selected schools is much smaller than in the number of inspections we can use in the first approach. The results for the inspections at the randomly selected schools generally show an improvement in school performance, but this improvement is usually not significant.

1 Introduction

A number of countries, mainly European ones, have governmental agencies to inspect schools. During these inspections, the inspectors evaluate the quality of the education the schools provide. The end product of an inspection typically includes a set of recommendations designed to improve the schools. Do these inspections succeed in improving the educational achievement of primary school pupils?

This paper focuses on the effect of school inspections on test scores of pupils in Dutch primary education. The Dutch Inspectorate of Education carries out two types of inspections which differ in intensity. The less intensive version (RST) takes approximately one day and the more intensive version (IST) takes two to three days. Estimating the effect of school inspections on primary school performance is difficult, because inspectors may not randomly select which schools they inspect. School inspectors are likely to visit poorly performing schools more often. They may also inspect schools for reasons a researcher can not observe. This non-random selection can produce an endogeneity bias in the estimates of the impact of school inspections. As a result, an estimated effect could actually be due to correlation between unobserved heterogeneity in the quality of schools and the inspectors' decisions about which schools to inspect.

We use two approaches to overcome this bias. First, we estimate models that include fixed school effects using test scores over the period 1999 to 2003. These models control for unobserved school factors that are time invariant. This ensures that correlation between the inspections and constant factors, such as the physical infrastructure of the school, will not bias the estimates. In this approach only changes in a school's quality over time can bias our estimates if they are correlated with (but not caused) by the school inspections. Our second approach also tries to rule out this possible bias. We exploit the fact that the Dutch Inspectorate of Education inspects a random sample of schools in order to compile their annual report on the state of Dutch education. In doing so, the inspectorate performs what is essentially a controlled experiment in which the sample of schools that are inspected represents a random experimental group. The advantage of this sample is that we can be more confident about the estimates. However, the number of inspections at randomly selected schools is much smaller than in the number of inspections we can use in the first approach.

Our study is related to research on accountability systems for schools. Many recent papers investigate the effect of using public performance indicators in education, see for example Jacob and Levitt (2003) and Figlio and Lucasc (2004). School inspections not only make schools accountable but also aim to provide recommendations for school improvement. In the literature various outcomes of school inspections have been studied, for instance Brimblecombe et al. (1996) have looked at the effect on teaching strategies, or Chapman (2001) who have studied the effect on changes in school policy. The literature on the effect of inspections visits on educational achievement is small and limited to studies for the UK. Three studies (Cullingford et al. (1999), Wilcox and Gray (1996), and Shaw et al. (2003)) investigate the effect of visits by the English Inspectorate, Ofsted, on examination results at secondary schools. All three studies find that Ofsted visits have a negative effect in the short term.The researchers suggest that this effect might come from stress and the need to prepare thoroughly for inspection visits. It should be noted that these studies control for observable differences between schools that are inspected and those that are not. This approach has the disadvantage that the estimated effects can be biased by unobserved differences between schools. In addition, these studies only estimated the effect of school inspections one year after an inspection. Our study is most related to a recent study by Rosenthal (2004). Using panel data on schools in British secondary education he also finds that Ofsted-inspections actually result in a decrease in the standardized exam scores of the students at the inspected schools. Our study extends previous research by using two approaches of which one is based on a natural experiment. Moreover, we are able to estimate the effect of inspections up to four years after an inspection.

Our main finding with the first approach, using a fixed effects model, is that school inspections lead to better performance of schools. In the first two years following an inspection, test scores increase by 2% to 3% of a standard deviation. Our analysis indicates that the more intensive inspections produce larger improvements in school performance than the less intensive ones. The improvement in Dutch elementary schools is strongest in the area of arithmetic and persists over the four years following an inspection that our data covers. For the three other subject areas covered by the Cito test and for the test score total, the improvement is significant in the two years following an inspection. Thereafter the estimated effects are typically positive and of similar magnitude to those in the first two years, but not significant.

With the second approach, based on the random sample drawn by the Dutch Inspectorate of Education, we only find significant effects on the arithmetic test. For the other test components, the estimates are not significantly different from zero. These estimates are, however, consistent with the overall results based on all the school inspections. The estimates are positive in the first two years following an inspection, and all estimates are positive and larger for the arithmetic component than for the other components.

In the following section, we discuss school inspections as they are performed in the Netherlands in more detail. Then, before presenting the results of our analysis, we will discuss the data we use and how we model this data to obtain estimates of the effects of an inspection. We conclude with a discussion of the findings.

2 School inspections in the Netherlands

One of the aims of the *Nederlandse Inspectie van het Onderwijs*, or the Dutch Inspectorate of Education, is to improve the quality of school education in the Netherlands. To achieve this, the inspectorate carries out inspections of primary schools. The school inspectors evaluate the quality of the education each school provides and recommends improvements. A school inspection usually takes two or three days and consist of questionnaires, observations of lessons and pre-structured interviews with the principal, teachers, parents, and pupils. At the end of an inspection, schools receive a report detailing its strengths and weaknesses. The inspection reports are also published on the Internet, including a table with the central assessment results. This is intended to make schools publicly accountable.

Since 1998, all primary schools have been inspected with a new and standardized evaluation instrument. This instrument is similar to the instrument of other inspectorates. It has been designed to assess primary schools based on student results and aspects of the educational process such as the competencies of teachers, the learning time, the pedagogical climate, and the management. Each school is assessed by a standard set of indicators. These consist of questionnaires, observation instruments, and pre-structured interviews.

Dutch law stipulates two statutory tasks for the inspectorate. One of these is to inspect schools. The other is to improve the quality of Dutch schools. The inspectorate aims to achieve this latter goal via the inspections themselves, as well as through the public report, the school quality card which follow an inspection. The assessment results and reports are designed to improve school performance directly by stimulating changes in policy and teaching methods and materials. An additional effect might come from follow-up activities of other parties involved, activities often directed by the assessment results. Examples are special measures for poorly performing schools and extra funding, support or control by local governments or school boards.

The treatment we are studying consists of the standardized school inspections, the public assessment report and the follow-up activities. The inspections and assessment reports are identical for every school, although there are differences in the assessment results. The follow-up activities vary among schools and are part of the treatment in our study.

In the years 1998-2003, there were two types of inspections. The RST inspection was a shorter version, and the IST, a more extensive one. Both versions were similar, but the more extensive IST was based on more measures of educational quality. In general, the inspectorate used the regular assessment instrument for all schools, while the extensive method was intended to be used as an instrument for follow-up inspections at under-performing schools.[1] The IST inspections were also performed on a random sample of schools. This sample has been drawn by

[1] However, in our analysis of the data, we were unable to find evidence that the IST inspections we performed as a follow-up at under-performing schools.

the Inspectorate to compile the annual report on the state of Dutch education. We denote these inspections as RIST. Table 2.1 lists the number of inspections of each type by year in our data. Dutch primary education has approximately 7600 schools.

Table 2.1 Number of Inspections in Sample

Year[a]	RIST	IST[b]	All IST[c]	RST
1998	-	643	643	-
1999	370	285	456	1406
2000	181	171	369	1209
2001	199	84	305	825
2002	-	90	245	607

[a] RIST inspections are dated according to the school year in which the Dutch Inspectorate selected which schools were to receive a RIST inspection. This selection was performed at the beginning of the school year. IST and RST inspections are dated according to the calendar year in which the inspection report was completed.

[b] The reported number excludes all RIST inspections for which a report was completed during the calendar year.

[c] In any given year, the number of inspections for the RIST and the IST will not added up to the number of inspections in the 'All IST' column. This is due to the difference between the school year dating of the RIST and the calendar year dating of the IST.

The number of inspections listed in the table for the IST and RST inspections are dated after the inspections took place. These dates indicate the calendar year in which the inspection report, which followed the actual inspection, was completed. The number of IST and RIST inspections do not add up to the total number of IST inspections due to the difference between the school year dating of the RIST and the calendar year dating of the IST inspections.

According to the Dutch Inspectorate of Education, the selection of schools in the years 1998 and 2002 for the RIST inspections was not entirely random. We do not, therefore, separate the RIST from the set of all IST inspections for these years. The RIST inspections in the years 1999, 2000, and 2001 were performed at randomly selected schools according to the Dutch Inspectorate. We have performed logit regressions of these inspections on student and school characteristics and the CITO test scores to confirm that these inspections were in fact carried out at randomly selected schools. We did not find systematic differences between schools in the random sample and schools that were not selected in this sample.

Finally, we note that the schools inspections are not uniformly distributed throughout the sample period. There were in fact more inspections of all three types in 1999 than in the following years. The set of all IST inspections and the RST inspections show a steady decline in numbers over time. For this reason, we have opted to include year dummy variables in our models of the test scores.

3 Data

The dependent variable in our analysis is the students' score on a standardized multiple-choice test called the Cito test. We use five years of test scores for the Cito test. More than 80% of primary schools administer this test to pupils in the final year of their primary education, which is called group 8. The average age of these pupils is 12 years. 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. There are five components to the test. These five components include the four listed above, only reading is tested in a separate component, and is not tested in the language component. The Cito test score total is made up of four of the five components, with world orientation being left out. This is due to the fact that world orientation is optional.

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 also currently used to judge the quality of primary schools. Parents use this information when choosing a primary school for their children. Every year the test received considerable media attention, with national newspapers and television reporting on the most recent results.

In this paper we use the test scores of all pupils in group 8 tested in 1999, 2000, 2001, 2002 and 2003. Our sample consists of approximately 720 000 pupils in 6 230 schools. The data set includes the standardized total Cito test score as well as the component scores. We drop those standardized scores which are lower than -2.66, because we believe that many of these students have special needs and in some primary schools are excluded from taking the Cito test. As a result of this corrective measure, we exclude 3 973 test scores from our analysis.[2]

We also exclude the world orientation test component from our analysis. Many schools do not administer this component, because it is optional. As a result these component scores are less likely to adequately reflect the general school population.

[2] Results obtained without excluding these 3 973 test scores indicate that our results are not effected by their exclusion.

Our main explanatory variables are the school inspections. As discussed in the previous section, there are two types of inspections: a shorter version (RST) and a longer version (IST). The IST inspection has also been performed on a random sample of schools (RIST). At the individual level, we have information on the gender of the student. At the school level, we have information about the number of students from each school who took part in the Cito test in each year. This provides a measure of the school size. Also available at the school level in each year are the shares of the school student body that fall into each one of five categories created by law and used to determine the level funding of each school by the government. Each category is determined by the socioeconomic background of the parents and is used to determine the amount of money a school receives for a certain pupil via a weighting factor. For example, a school receives 25% more funds for pupils of poorly educated Dutch parents. The government allocates 90% more funds for the children of poorly educated parents from an ethnic minority. The share of pupils belonging to these categories allows us to control for demographic effects that influence the overall performance potential of each school.

Table 3.1 lists the means and standard deviations of the variables we use in our analysis. The first column reports the means for the entire sample. The slight increase in the score means and decline in the standard deviations is caused by the exclusion of the lower tail implied by dropping pupils with CITO test score totals below -2.66. The three remaining columns correspond to the means obtained for schools subject to a RST, an IST, and a RIST inspection. The means are calculated using the data from the year in which the inspections took place. The column for the IST inspections represents the means obtained for both the randomly and non-randomly selected schools.

The results indicate that the measured characteristics of each group of schools that received one of the three inspection types closely mirror those found for the entire sample. The lower test score means for all IST inspections is, however, worth mentioning. This is an indication that inspectors tended to visit relatively poorly performing schools more often with the IST-instrument.

Table 3.1 Means for entire sample and randomly inspected schools, 1999-2003

Variable	School Group			
	All	RST	All IST	RIST
Cito test score				
Total	0.02	0.01	− 0.04	− 0.01
(standard deviation)	(0.98)	(0.98)	(0.99)	(0.98)
Arithmetic	0.01	0.01	− 0.04	− 0.01
(standard deviation)	(0.99)	(0.99)	(1.00)	(0.99)
Language	0.01	0.01	− 0.04	− 0.01
(standard deviation)	(0.98)	(0.98)	(0.99)	(0.99)
Information	0.02	0.01	− 0.04	− 0.01
(standard deviation)	(0.98)	(0.98)	(1.00)	(0.99)
Reading	0.01	0.01	− 0.03	− 0.01
(standard deviation)	(0.98)	(0.98)	(1.00)	(0.99)
Socioeconomic index				
1 (least disadvantaged)	0.12	0.12	0.12	0.11
2	0.29	0.26	0.25	0.27
3	0.32	0.34	0.32	0.32
4	0.11	0.13	0.12	0.13
5	0.08	0.08	0.09	0.09
6	0.04	0.04	0.04	0.04
7 (most disadvantaged)	0.04	0.04	0.06	0.04
School denomination				
Public	0.33	0.32	0.36	0.35
Catholic	0.32	0.32	0.28	0.30
Protestant	0.29	0.30	0.30	0.30
Montessori/Dalton	0.05	0.05	0.05	0.05
Other	0.01	0.01	0.01	0.01
Urbanization school area				
Very High	0.12	0.12	0.15	0.13
High	0.17	0.17	0.19	0.18
Modest	0.19	0.18	0.19	0.20
Low	0.28	0.28	0.26	0.27
Rural	0.24	0.25	0.20	0.21
Funding weights				
1.0	0.719	0.709	0.693	0.703
1.25	0.153	0.166	0.158	0.162
1.4	0.001	0.001	0.001	0.001
1.7	0.002	0.003	0.002	0.002
1.9	0.124	0.121	0.146	0.132
School size	223.8	217.5	227.1	224.0
Number Cito per school	24.83	24.99	25.35	24.61
% girls	0.49	0.50	0.50	0.49
% missing gender	0.02	0.02	0.02	0.01
Sample size $= N$	716010	105956	37378	20135
School measurements	28518	4437	1493	816

4 Empirical strategy

To determine the effect of school inspections on primary schools it would be convenient if the inspected schools were randomly selected. Unfortunately this is not likely to have been the case. It is therefore probable that the inspections and the unobserved heterogeneity in the test scores are correlated. For example, it may well be the case that school inspections occurred predominately at poorly performing schools. If this is the case, then school quality and the inspections will be negatively correlated. As a result, a simple regression of Cito test scores on the set of dummy variables covering the history of school inspections will produce negatively biased estimates.

We use two approaches to overcome this type of bias. The first approach is to use a standard fixed effects model. The second approach exploits a random sample created by the Dutch Inspectorate of Education for compiling the annual report on the state of Dutch education. In both approaches, we essentially compare the improvement in test scores following a certain type of inspection with the change in test scores over the same period at those schools where this type of inspection did not take place. In the first approach, we focus on both types of inspections. In the second type, we compare the performance of schools in the random sample with the performance of the other schools.

Analyzing all school inspections using a fixed effects model is feasible, because we have pupil test scores measuring school performance before the inspections took place. Our second approach, based on the inspections at randomly selected schools, does not require this data. The disadvantage of this latter approach, however, is that there are fewer inspections to analyze. As a result, there is more uncertainty about the estimates based on the random sample of inspections.

4.1 First approach

The first approach, using a standard fixed effects model, assumes that there is only correlation between the inspections and the time-invariant components of the unobserved heterogeneity. Provided school quality remains essentially constant over the course of the sample period, the fixed effects method will eliminate the problem of bias due to variation in schools quality.

We base our analysis on univariate models for the four test score components and the Cito test total. These models are of the following form:

$$y_{ijt} = I'_{jt}\beta + X'_{ijt}\delta + v_j + \varepsilon_{ijt}, \tag{4.1}$$

Here, y_{ijt} is the Cito score for one the five components of the Cito test or the test score total over time, where $i = 1,\dots,N_{jt}$ is the index used to designate the individual student, $j = 1,\dots,S$ denotes the school, and $t = 1,\dots,T$ time, which in this case is the year. The test score is assumed to be a linear function of the treatment variables, the control variables, and the two residual terms $v_j + \varepsilon_{ijt}$. The vector I_{jt} consists of the treatment variables, while the vector X_{ijt} is the set of explanatory variables. The effect of the inspections is given by the parameter vector β. The vector δ corresponds to the vector of effects of the control variables. The residual consist of a school effect v_j, and the disturbance term ε_{ijt}. The disturbance term ε_{ijt} is assumed to be independently and identically distributed with mean zero and variance σ_ε^2.

The treatment variable vector I_{jt} is made up of variables derived from the primary school inspection history. The term $I'_{jt}\beta$ in (4.1) can be written out as follows.

$$I'_{jt}\beta = I_{1jt}\beta_1 + I_{2jt}\beta_2 + I_{3jt}\beta_3 + I_{4jt}\beta_4. \tag{4.2}$$

The four variables I_{1jt}, I_{2jt}, I_{3jt}, and I_{4jt} on the right hand side are dummy variables. The variable I_{1jt} indicates whether or not there was an inspection at school j a year ago (in the year $t-1$). For example if y_{ijt} represents the Cito test score administered in February 2001, then $t = 2001$. In this case a value of $I_{1j2001} = 1$ indicates that the school j was inspected sometime during the year 2000. Similarly, the variables I_{2jt}, I_{3jt}, I_{4jt} indicate whether or not there was a random inspection at school j two years ago, three years ago, and four years ago, respectively. In the definition of these variables, we make no distinction between the various types of inspections. This means we assume that the RST and IST both have the same effect on school performance.

In order to be able to obtain unbiased estimates for the elements of the parameter vector β in (4.1) using the fixed effects method, the treatment vector I_{jt} and the residual term ε_{ijt} must be uncorrelated: $cov(I_{jt}, \varepsilon_{ijt}) = 0$. In the case of the non-random IST and the RST inspections, we do not know how the inspection service selected which schools were to be inspected. We cannot therefore rule out the possibility that inspectors tended to visit certain types of schools more than others. If inspectors were more likely to visit weaker schools, then the correlation between I_{jt}, and the school effect v_j would result in a negative bias in the estimated values.

Note that our data does not permit us to follow individuals over time as is typically the case in a panel. Each year the students in the data set leave primary school and therefore do not reappear the following year. As a result the school effect v_j is the only source of serial correlation in the test scores.

We also use a second fixed effects model which we obtain by relaxing the assumption that the IST and RST inspections have the same effect. This leads to model (4.3).

$$y_{ijt} = RST'_{jt}\hat{\beta} + I\hat{S}T'_{jt}\tilde{\beta} + X'_{ijt}\delta + v_j + \varepsilon_{ijt} \tag{4.3}$$

Here we have that

$$RST'_{jt}\beta = RST_{1jt}\hat{\beta}_1 + RST_{2jt}\hat{\beta}_2 + RST_{3jt}\hat{\beta}_3 + RST_{4jt}\hat{\beta}_4 \tag{4.4}$$

and

$$I\hat{S}T'_{jt}\beta = I\hat{S}T_{1jt}\tilde{\beta}_1 + I\hat{S}T_{2jt}\tilde{\beta}_2 + I\hat{S}T_{3jt}\tilde{\beta}_3 + I\hat{S}T_{4jt}\tilde{\beta}_4. \tag{4.5}$$

The dummy variable RST_{kjt}, $k = 1, \ldots, 4$ indicates whether school j in the year t had an RST inspection k years ago. $I\hat{S}T_{kjt}$ is defined similarly for IST inspections (at both the randomly selected and the non-randomly selected schools).

We note that all versions of our model include dummy control variables for the IST inspections carried out in 1998. We have no test score available before these interventions took place. For this reason, the effect of these inspections after one year (in 1999) ends up in the fixed effect term v_j of those schools which were inspected in 1998. We therefore include control dummy variable for the IST inspections from 1998 for the additional effect of these inspections after two, three, four, and five years. This means that we are unable to measure the total effect of these inspections on school performance. We have therefore opted not to report them.

4.2 Second approach

The second approach we use to overcome the potential problem of selection bias is based on a random sample created by the Inspectorate of Education for compiling the annual report of the state of Dutch education. This random sample provides an experimental group of schools at which school inspections are held. All other schools form the control group. We compare the outcomes of schools in the experimental and control groups in the years following the school inspections. By creating a random sample of schools to be inspected, the Dutch Inspectorate of education is actually performing a controlled experiment.

To be able to obtain estimates based on the IST inspections at randomly selected schools, we define dummy variables for all three types of inspections. To do this we split the dummy variable $I\hat{S}T_{jt}$ into the variables IST_{jt} and $RIST_{jt}$. This model version is given in (4.6).

$$y_{ijt} = RST'_{jt}\hat{\beta} + IST'_{jt}\tilde{\beta} + RIST'_{jt}\beta^{\dagger} + X'_{ijt}\delta + v_j + \varepsilon_{ijt}. \tag{4.6}$$

Here the dummy variable IST_{jt} represents the inspection history of the IST inspections at the non-randomly selected schools, and the variable $RIST_{jt}$ is based on the history of the IST inspections at the randomly selected schools. These dummy variables are defined in the same manner as in (4.2). There are, however, only three dummy variables in the vector $RIST_{jt}$, because we have no information on these inspections for the year 2002, see table 2.1.

By specifying a fixed effects model in terms of the intervention variables RST_{jt}, IST_{jt}, and $RIST_{jt}$, we can not only check whether the RST and IST inspections produce similar improvements in school performance, but we can also check for the presence of bias in our estimates. If the dummy variables RST_{jt} and IST_{jt} are correlated with the disturbance term ε_{ijt}, the estimates for the parameters $\hat{\beta}$ and $\tilde{\beta}$ will be biased. The estimates for the effect of school inspection based on the inspections performed at the randomly inspected schools, $\beta^{\dagger}$ should be free from this bias. The fact that the school were randomly selected ensures that $RIST_{jt}$ will be independent of ε_{ijt}.

The estimates of $\tilde{\beta}$ and $\beta^{\dagger}$ are both measures of the effect of IST inspections on school performance. It would therefore be reasonable to expect that these parameters be equal. If the estimates are different, this may be an indication that there is some correlation between the time varying heterogeneity, disturbance term ε_{ijt}, and the IST inspections at the non-randomly selected schools. In the following section, we present the estimation results we obtained for the fixed effects models presented above.

5 Results

We present the estimated effects of school inspections in the tables 5.1 to 5.5. In the presentation of our results, we adopt the notation of three asterisks to denote an estimate that is significant at the 0.1% level, two asterisks to denote an estimate that is significant at the 1% level, and one asterisk at the 5% level. The standard error is shown below each estimate in parentheses. The results are grouped by Cito test component, with one table for each component. We report only the estimates we obtained for the model with control variables. The control variables included are the school size, socioeconomic index, funding weights, and number of cito test scores administered, as well as the ratio of the number of test scores administered to the school size, and gender. For details on these variables see table 3.1. The results for the model without control variables are essentially the same. We have chosen not to reproduce the results for the model without control variables to avoid presenting too many tables.

In each table we list the estimated effects of the different types of inspections. The first row reports the combined effect of all school inspections bundled together. The estimates in this row are based on (4.1) in which we assume that the effect of the school inspections is the same regardless of the type of inspection. The second row shows the estimates for the RST inspections. These estimates are based on (4.3). In the last two rows we report the estimates for the IST inspections. The third row is based on all IST inspections performed at both the randomly, as well as the non-randomly selected schools in (4.3). The last row lists the estimates for the randomly selected IST, or the RIST inspections. These results were obtained using (4.6).

The first table 5.1 lists the estimation results we obtained based on the Cito test total score. The estimates of the effects of the RST inspections, $\hat{\beta}$, as well as those for all IST inspections, $\tilde{\beta}$, are highly significant in the two years following an these inspections. This is also true of the estimates for the combined effect of all inspections β. Although the estimates obtained for the IST inspections at randomly selected schools, $\beta^{\dagger}$ are not significantly different from zero, they are positive in the first two years following an inspection. It is important to note that the standard errors for the estimates of $\beta^{\dagger}$ are nearly double those for β. This is due to the smaller number of inspections at randomly selected schools. This makes it more difficult to find an effect that is significantly different from zero.

An important feature of these results is that all our estimates are positive. The mean squared error of our estimates increases with the elapsed time since an inspection. This is due to the decreased number of inspections that are available to identify these estimates. In the case of $\hat{\beta}_4$ and $\tilde{\beta}_4$ these estimates are only based on the inspections in the year 1999. This contrasts with the estimates for $\hat{\beta}_1$ and $\tilde{\beta}_1$, which are based on the inspections in the four years over the period of 1999 to 2002. This increase in the standard error as the time elapsed since an inspection grows, means that it is increasingly difficult to accurately measure the effect of an inspection in our sample as the time elapsed since an inspection grows. We can not rule out the possibility therefore that there is a positive effect in each year following an inspection, but that we are unable to measure this effect with sufficient precision to demonstrate this. This hypothesis is supported by the fact that the estimates of $\tilde{\beta}$ are also significant in the fourth year.

We can also see from the tables that the estimates of $\tilde{\beta}$ are larger than those for $\hat{\beta}$. This suggests that the more intensive IST inspections lead to a larger improvement in test score than the shorter RST inspections.

As is typical for all our estimates, regardless of which test component they are based on, the model specification does not seem to significantly effect the estimates we obtain. This suggests that our results are robust to changes in model specification.

Table 5.1 Estimated Effect of School Inspections on Cito Test Total[a]

Type of Inspection	Time elapsed since inspection			
	1 year	2 years	3 years	4 years
	β_1	β_2	β_3	β_4
All inspections (β in 4.1)	0.019^{***}	0.023^{***}	0.009	0.013
	(0.004)	(0.006)	(0.008)	(0.010)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
RST inspections ($\hat{\beta}$ in 4.3)	0.017^{***}	0.025^{***}	0.009	0.007
	(0.004)	(0.006)	(0.008)	(0.010)
	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$
All IST inspections ($\tilde{\beta}$ in 4.3)	0.028^{***}	0.023^{**}	0.016	0.039^{**}
	(0.006)	(0.008)	(0.010)	(0.013)
	$\beta_1^\dagger$	$\beta_2^\dagger$	$\beta_3^\dagger$	
Random IST inspections ($\beta^\dagger$ in 4.6)	0.005	0.006	0.000	
	(0.008)	(0.010)	(0.013)	

[a] The control variables are the school size, socioeconomic index, funding weights, and number of cito test scores administered, and the ratio of the number of test scores to the school size, and gender.

The estimates based on the arithmetic test component show the largest increases in school performance following an inspection. Table 5.2 gives these estimates. The estimates of the effect of all inspections, β, of the RST inspections, $\hat{\beta}$, and of all IST inspections $\tilde{\beta}$ are significantly different from zero in all years. The estimates for the effect from the RIST inspections, $\beta^\dagger$ are not generally significantly different from zero, although $\beta_2^\dagger$ is. The $\beta^\dagger$ are, however, all positive. Once again, it is also the case that the standard errors of the $\beta^\dagger$ are also higher, indicating that it is more difficult to accurately measure the effect of these inspections, because there are fewer of them.

Table 5.3 lists the estimates based on the language test component. These estimates show improvements of all the test components in the first two years following an inspection. The table for the estimates based on the information test component, table 5.4, as well as the one based on

Table 5.2 Estimated Effect of School Inspections on Cito Test Arithmetic[a]

Type of Inspection	Time elapsed since inspection			
	1 year	2 years	3 years	4 years
	β_1	β_2	β_3	β_4
All inspections (β in 4.1)	0.027***	0.036***	0.028***	0.033***
	(0.004)	(0.006)	(0.008)	(0.010)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
RST inspections ($\hat{\beta}$ in 4.3)	0.025***	0.034***	0.026**	0.025*
	(0.005)	(0.006)	(0.008)	(0.010)
	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$
All IST inspections ($\tilde{\beta}$ in 4.3)	0.037***	0.043***	0.038***	0.060***
	(0.006)	(0.008)	(0.010)	(0.013)
	$\beta_1^\dagger$	$\beta_2^\dagger$	$\beta_3^\dagger$	
Random IST inspections ($\beta^\dagger$ in 4.6)	0.007	0.021*	0.021	
	(0.008)	(0.010)	(0.013)	

[a] The control variables are the school size, socioeconomic index, funding weights, and number of cito test scores administered, and the ratio of the number of test scores to the school size, and gender.

Table 5.3 Estimated Effect of School Inspections on Cito Test Language[a]

Type of Inspection	Time elapsed since inspection			
	1 year	2 years	3 years	4 years
	β_1	β_2	β_3	β_4
All inspections (β in 4.1)	0.011**	0.011*	− 0.005	− 0.005
	(0.004)	(0.006)	(0.008)	(0.010)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
RST inspections ($\hat{\beta}$ in 4.3)	0.009*	0.015*	− 0.004	− 0.009
	(0.004)	(0.006)	(0.008)	(0.010)
	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$
All IST inspections ($\tilde{\beta}$ in 4.3)	0.020***	0.005	− 0.002	0.015
	(0.006)	(0.008)	(0.010)	(0.013)
	$\beta_1^\dagger$	$\beta_2^\dagger$	$\beta_3^\dagger$	
Random IST inspections ($\beta^\dagger$ in 4.6)	0.005	0.000	− 0.012	
	(0.008)	(0.010)	(0.013)	

[a] The control variables are the school size, socioeconomic index, funding weights, and number of cito test scores administered, and the ratio of the number of test scores to the school size, and gender.

the reading test component, table 5.5, show a largely similar pattern. The estimates for the information component and the language component tend to be smaller, while those for the reading component tend to be somewhat larger. There is, in fact, only one negative estimate in the table for the reading component, whereas the table for the information component contains five negative estimates.

Table 5.4 Estimated Effect of School Inspections on Cito Test Information[a]

Type of Inspection	Time elapsed since inspection			
	1 year	2 years	3 years	4 years
	β_1	β_2	β_3	β_4
All inspections (β in 4.1)	0.012^{**}	0.014^{*}	-0.002	0.005
	(0.004)	(0.006)	(0.008)	(0.010)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
RST inspections ($\hat{\beta}$ in 4.3)	0.011^{*}	0.017^{**}	-0.001	0.001
	(0.005)	(0.006)	(0.008)	(0.010)
	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$
All IST inspections ($\tilde{\beta}$ in 4.3)	0.017^{**}	0.010	0.001	0.025
	(0.006)	(0.008)	(0.010)	(0.013)
	$\beta_1^\dagger$	$\beta_2^\dagger$	$\beta_3^\dagger$	
Random IST inspections ($\beta^\dagger$ in 4.6)	-0.004	-0.014	-0.019	
	(0.008)	(0.010)	(0.013)	

[a] The control variables are the school size, socioeconomic index, funding weights, and number of cito test scores administered, and the ratio of the number of test scores to the school size, and gender.

Table 5.5 Estimated Effect of School Inspections on Cito Test Reading[a]

Type of Inspection	Time elapsed since inspection			
	1 year	2 years	3 years	4 years
	β_1	β_2	β_3	β_4
All inspections (β in 4.1)	0.013^{**}	0.016^{**}	0.001	0.013
	(0.004)	(0.006)	(0.008)	(0.010)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
RST inspections ($\hat{\beta}$ in 4.3)	0.013^{**}	0.018^{**}	0.000	0.007
	(0.005)	(0.006)	(0.008)	(0.011)
	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$
All IST inspections ($\tilde{\beta}$ in 4.3)	0.017^{**}	0.017^{*}	0.008	0.039^{**}
	(0.006)	(0.008)	(0.010)	(0.013)
	$\beta_1^\dagger$	$\beta_2^\dagger$	$\beta_3^\dagger$	
Random IST inspections ($\beta^\dagger$ in 4.6)	0.010	0.002	0.004	
	(0.008)	(0.011)	(0.013)	

[a] The control variables are the school size, socioeconomic index, funding weights, and number of cito test scores administered, and the ratio of the number of test scores to the school size, and gender.

In general, the results follow a number of patterns. Firstly, the estimates for the IST inspections, $\tilde{\beta}$ are the largest, and those for the IST inspections at randomly selected schools, $\beta^{\dagger}$ are the smallest. Those for the RST, $\hat{\beta}$, and for all inspections, β, lie in between.

Secondly, the measured effects three and four year after the inspections are typically not significantly different from zero, and, particularly for the language and information components, can be negative. However, all estimates also follow a third pattern: that of increasing mean squared errors as time elapses following an inspection. We can therefore not rule out the possibility that the inspections do produce permanent improvements in school performance, but that we do not have a sufficiently long sample period to be able to demonstrate this. It is worth noting, for example, that there is not a single negative estimate in the tables to be found that is also significantly different from zero.

In general, the tables make clear that the most dramatic improvements in test scores following an inspection are to be found for the arithmetic test component. These improvements are positive and significantly different from zero for all four years or all parameter estimates with the exception of two of the three values for $\beta^{\dagger}$.

6 Conclusions

In this paper we investigate whether inspections by the Dutch Inspectorate of Education lead to an improvement of test scores. It should be noted that we did not investigate the effect of the existence of the Dutch Inspectorate of Education. In the Dutch education system any school can be visited by schools inspectors and the threat of an inspection may have an impact on school performance. Our analysis only focuses on the effect of the school visits by inspectors and their follow-up activities. To avoid selection bias by inspectors choosing schools to visit, we use two approaches for estimating the effects of school inspections. The first approach is to use a standard fixed effects model. The second approach exploits a sample of randomly selected schools originally drawn for the purpose of compiling the annual report of the state of Dutch education.

Our main finding with the first approach is that school inspections lead to better performance of schools. In the first two years following an inspection test scores increase by 2% to 3% of a standard deviation. Our analysis also indicates that the more intensive inspections produce larger improvements in school performance than the less intensive ones.

The improvement in Dutch elementary schools is strongest in the area of arithmetic and persists over the four years following an inspection. For the three other subject areas covered by the Cito test and for the test score total, the improvement is significant in the two years following an inspection. Thereafter, the estimated effects are typically positive and of similar magnitude to those in the first two years, but not significant.

For the second approach based on the random sample drawn by the Dutch Inspectorate of Education, we only find significant effects for the arithmetic component of the Cito test. However, the estimates for the other test components are consistent with the overall results based on all the school inspections. The estimates are positive in the first two years following an inspection, and all estimates are positive and larger for the arithmetic component than for the other components. The small number of inspections in the random sample reduces the statistical power in the second approach which may explain the insignificant results.

Why does the first approach yield larger results than the second approach? One possible explanation is be that the nonrandom inspections might more often be targeted at schools with greater potential for improvement. In this case, we would expect that there are more schools that do not benefit from the recommendations of the Inspectorate in the random sample. Nonrandomly selected schools would then on average show more improvement than the randomly chosen ones in the second approach. This is related to the distinction made in the evaluation literature between average treatment effects and average treatment effects on the treated, see for instance Cameron and Trivedi (2005). The difference between the first and second approach could be caused by the difference between these two treatment effects.

We conclude that both approaches indicate positive effects of school inspections on achievements of pupils in primary education. If school visits of two to three days improve test scores by 2% to 3% of a standard deviation, this seems a very cost-effective intervention compared to other interventions. For instance, in the famous Star experiment in Tennessee a class size reduction of seven pupils for four years increased average test scores between 10% and 20% of a standard deviation. Although the benefits of the class size intervention are larger, the cost is also likely to be greater. Reducing class sizes by seven pupils involves increasing the teacher labor force by approximately one third for four years; the cost of a school inspection and report by inspectors is a fraction of this cost. Our estimates in the second approach are smaller than 2% to 3%, but even with very small improvements of test scores the benefit-cost ratio of school inspections compares favorably to class size reduction.

References

Brimblecombe, N., M. Shaw and M. Ormston, 1996, Teachers’ intention to change practice as a result of Ofsted school inspections, *Education Management and Administration*, vol. 24, no. 4, pp. 339–354.

Cameron, A.C. and P.K. Trivedi, 2005, *Microeconometrics: Methods and Applications*, Cambridge University Press, New York, USA.

Chapman, C., 2001, Changing classrooms through inspections, *School Leadership and Management*, vol. 21, no. 1, pp. 59–73.

Cullingford, C.I., S. Daniels and J. Brown, 1999, The effects of Ofsted inspection on school performance, *School Leadership and Management*, vol. 19, no. 4, pp. 323–526.

Figlio, D.N. and M.E. Lucasc, 2004, Do high grading standards affect student performance?, *Journal of Public Economics*, vol. 88, no. 9-10, pp. 1815–1834.

Jacob, B.A. and S.D. Levitt, 2003, Rotten apples: An investigation of the prevalence and predictors of teacher cheating, *The Quarterly Journal of Economics*, vol. 118, no. 3, pp. 843–877.

Rosenthal, L., 2004, Do school inspections improve school quality? Ofsted inspections and school examination results in the UK, *Economics of Education Review*, vol. 23, no. 23, pp. 143–151.

Shaw, I., D.P. Newton, M. Aitkin and R. Darnell, 2003, Do Ofsted inspections of secondary schools make a difference to GCSE results?, *British Educational Research Journal*, vol. 29, no. 1, pp. 63–75.

Wilcox, B. and J. Gray, 1996, *Inspecting Schools: Holding Schools to Account and Helping Schools to Improve*, Open University Press, Buckingham.