

Mechanisms of change in Dutch inspected schools; comparing schools in different inspection treatments

M. Ehren

N. Shackleton

British Journal of Educational Studies

The DOI of your paper is: 10.1080/00071005.2015.1019413

Abstract

The need for education systems and schools to improve and innovate has become central to the education policy of governments across the world according to Gray (2014). School inspections are expected to play an important role in promoting such continuous improvement and to help schools and education systems more generally to consider the need for change and improvement. This paper aims to enhance our understanding of the connections between school inspections and their impact on school improvement, using a longitudinal survey to principals and teachers in primary and secondary education. Random effects models and a longitudinal path model suggest that school inspections particularly have an impact on principals, and not so much teachers. The results indicate that the actual impact on improved school and teaching conditions, and ultimately student achievement is limited. Schools in different inspection categories report of different mechanisms of potential impact; the lack of any correlation between accepting feedback, setting expectations and stakeholder sensitivity and improvement actions in the schools suggests that impact of school inspections is not a linear process, but operates through diffuse and cyclical processes of change.

Introduction

The need for education systems and schools to improve and innovate has become central to the education policy of governments across the world according to Gray (2014). The Bratislava memorandum of SICI, the European Association of Inspectorates of Education, specifies that school inspections are expected to play an important role in promoting such continuous improvement and to help schools and education systems more generally to consider the need for change and improvement¹. As the memorandum outlines, the relationship between inspection and innovation is complex and the role of inspections in raising quality and standards of achievement has been discussed on many occasions.

Recently, a number of comparative studies such as *Governing by Inspection* (Grek et al., 2013), the *Impact of School Inspections on Teaching and Learning* project (Ehren et al., 2013) and research by the OECD (Looney, 2009; OECD, 2013) and SICI (Gray, 2014) have been carried out to obtain a better understanding of the effectiveness of school inspection and its contribution to educational improvement in a number of different contexts and education systems. These and other studies (see Luginbuhl et al, 2009; Hussain, 2012; Allen and Burgess, 2012) indicate that school inspections can have a great impact on what students learn and how they learn. The standards Inspectorates use to assess educational quality and teaching and learning in schools during inspection visits, the sanctions for failing and underperforming schools and the rewards for highly effective and well-functioning schools stimulate and pressure schools to meet nationally defined targets and objectives. School inspections may however also lead to unintended negative consequences for teaching and learning in schools. Possible negative consequences have been categorized by De Wolf and Janssens (2007) as intended and unintended strategic behaviour of schools and teachers. These types of behaviours may negatively affect student achievement in schools. Rosenthal (2004) for example found a decrease in examination results of pupils in England in secondary education in the year of the inspection visit, and Shaw et al (2003) found that inspection did not improve examination achievement in maintained comprehensive schools. Other studies, such as by Matthews & Sammons (2004) who analysed results for all secondary schools in the year before the visit and two years

¹ <http://www.sici-inspectorates.eu/About-us/Vision-mission/The-Bratislava-Memorandum-is-on-the-Website>

following on from the inspection however did not indicate such a dip in student achievement after inspections.

These contradictory findings portray complex and varied links amongst elements of inspections, mechanisms of impact and school outcomes that make the study of the functioning of inspection systems complicated. Understanding the effects and the effectiveness of school inspections across different contexts entails detailed analysis of diverse contexts and schools systems, multiple system layers and interaction effects among approaches to school inspections that are currently implemented.

This paper aims to enhance our understanding of the connections between school inspections and their impact on school improvement, and how such connections may be different for teachers, principals and primary and secondary schools. An attempt is made to unpack how these connections change over time as schools implement feedback from recent inspection visits, or start to prepare for upcoming visits. The focus of the paper is on the Inspectorate of Education in the Netherlands, and attempts to answer two questions:

- (1) What impact (school improvement and unintended consequences) do Dutch school inspections have on primary and secondary schools, according to principals and teachers?
- (2) Which intermediate processes/mechanisms (these include providing feedback/setting expectations/stakeholders' sensitivity to inspection reports) explain this impact, according to principals and teachers in Dutch primary and secondary education?

The following section first explains how schools are inspected in the Netherlands. A brief literature review is then presented which reflects on the key mechanisms of impact of school inspections, and informs the theoretical framework of the study.

School inspections in the Netherlands²

The Dutch Inspectorate of Education, established in 1801, is one of the oldest operating Inspectorates in Europe. Its working methods, like those of other inspectorates, have evolved greatly over time, particularly in the last decade. At the time of the study (2010-2013), the Dutch Inspectorate of Education used a risk based inspection method to inspect schools in primary and secondary education

This method included annual early warning analyses of potential risks of failing educational quality in all schools. In these analyses, student achievement results on standardized tests, self evaluation reports and financial reports of schools, complaints of parents and news items in the media were all used to identify potentially failing schools. Students' results (corrected for their socio-economic backgrounds) on national standardized tests and examinations were the primary indicator in the analyses and were used to classify schools into one of three categories; schools in the 'green' category are considered to have no risks of failing, 'orange' schools have potential risks of failing, whereas 'red' schools have high risks of failing.

Schools in the 'green' inspection category received a 'basic' inspection treatment which meant there was no further inspection activity in the school that year. The Inspectorate considered student results to be a good predictor of the educational quality of schools and expects quality of the teaching and school organisation to be adequate and not in need of an external evaluation.

The Inspectorate of Education scheduled desk research on schools in the 'orange' category. These schools were requested to send in the student achievement results of students in intermediate grades in literacy

² The following section was taken from <authors>)

and mathematics. The Inspectorate also analysed additional documents about the school, such as annual reports. In cases in which this desk research showed no risks (the documents are in order and student achievement results in intermediate grades are sufficient and there are no indications of risk); the school was reassigned to the 'green' category. The school board however received an informal warning in case the achievement of students in the final grade was below average or was declining. An interview with the school board was scheduled in cases in which the desk research pointed to potential risks. Potential risks were discussed during this interview, as well as the capacity of the school board to address and solve these risks. An additional inspection visit to the potentially failing school was arranged in the event that this interview did not provide the Inspectorate of Education with sufficient information or in cases in which the capacity of the school board to address the risks was evaluated as inadequate. During this visit, the inspection framework was used to assess educational quality in the school as 'basic', 'weak' or 'very weak'.

The Inspectorate of Education also undertook desk research of schools in the 'red' category, comparable to the desk research of schools in the 'orange' category. School boards of the 'red' category schools were interviewed and they received a full inspection visit to evaluate their educational quality on nine indicators which covered students' results and educational processes in the school. During this visit, the inspection framework was again used to assess educational quality in the school as 'basic', 'weak' or 'very weak'.

The Inspectorate of Education instructed the school board of a weak school to formulate a plan of action aimed at improving quality. The Inspectorate tested the plan and laid down performance agreements in an inspection plan. This plan specified when the quality should be up to par again and what (interim) results the school must attain. It also specified the indicators the Inspectorate of Education would assess in (interim) inspection visits to the failing school. The school board had to commit to the inspection plan. Weak schools that did not improve within two years ended up in the regime imposed on very weak schools. These schools were scheduled for a meeting between the school board and the Inspectorate management and an official warning was given. If these activities did not yield the result agreed upon, the Inspectorate reported the school to the Minister, along with a proposal for putting sanctions in place. On the basis of this report, the Minister could proceed to impose administrative and/or financial sanctions.

Theoretical framework

Two recent reviews (Klerks, 2013; Nelson and Ehren, 2014) summarize the impact of school inspections on improvement of schools, schools' self-evaluations and ultimately student outcomes in maths and literacy, as well as unintended consequences of inspections. These reviews give a very mixed view on the ultimate outcomes of school inspections, and why they are causing changes in schools. The standards and thresholds used to assess schools during inspection visits and the sanctions and rewards deployed to improve schools seem to be the dominant aspects of school inspections affecting (both positive and negative) change in schools.

Standards and threshold

Hanushek and Raymond (2002) for example point to rational choice theory to describe how standards, the thresholds in performance targets and related sanctions and rewards may influence actions in schools. Standards present the details of what is expected of schools; they create boundaries or domains for attention with respect to educational quality. Some Inspectorates include a specific threshold or target in their framework to identify whether a school is failing or not. In the Netherlands, targets have been set to categorize schools as basic weak, or very weak. These standards and particularly the threshold to identify failing schools are expected to be important aspects of the impact of school inspections. Hanushek and Raymond (2002) describe how school officials in test-based accountability systems in the USA would select the action that they perceive to have the highest yield, given their planning horizon, budget and appetite for risk. These authors found failing schools in the USA on the brink of being sanctioned to make

dramatic improvements in the year after identification of these failures. Such improvements are likely to be highly strategic and focused on 'quick fix' measures. De Wolf and Janssens (2007) for example describe how schools concentrate on short term goals or focus on quantifiable phenomena in the inspection framework to meet standards in inspection frameworks, or refrain from innovating out of fear of not complying to the standards, particularly when the inspection standards remain the same for a long period of time, are used in a high stakes context and in an inflexible manner.

Sanctions and rewards

Some studies also suggest that sanctions and rewards have a positive effect on educational quality in schools. The operating assumption in these studies is that schools work harder to perform well when something valuable is to be gained or lost; information and feedback alone is seen as insufficient to motivate schools to perform to high standards (Malen, 1999; Elmore and Fuhrman, 2001; Nichols et al, 2006). Schools may receive rewards for good performance (such as financial bonuses or awards) or may be sanctioned when assessed to be failing. Sanctions are for example naming and shaming of the school on the internet, providing pupils in the school with vouchers to transfer to another school or fines.

Heubert and Hauser (1999) found a significant relationship between the level of incentives for schools and students and the extent to which the curriculum and teaching in schools change. Responses to inspection tend to be most focused and effective where funding is at stake or exposure is higher, according to Matthews and Sammons (2004). Formal sanctions, like forced reconstitution of consistently low performing schools, were more likely to promote responses than just embarrassment from grading schools and reporting results publicly. Sanctions raise awareness of the importance of the standards as well as force schools to comply with the standards.

High stakes (test-based) accountability systems have however also been known to produce harmful consequences (Heubert and Hauser, 1999; Koretz, 2003; Stecher, 2001). Sanctions and rewards may discourage desirable behaviour or may stimulate unintended and undesirable behaviour. Kerr (1975) describes how organisms seek information concerning what activities are rewarded, and then seek to do (or at least pretend to do) those things, often to the virtual exclusion of activities not rewarded. The extent to which this occurs depends, according to Kerr (1975), on the perceived attractiveness of the rewards offered. According to Elmore and Fuhrman (2001), schools operating under severe sanctions as reconstitution and probation or special measures do not appear to be making fundamental changes in their core processes. Instead they seem to exercise considerable emphasis on short term strategic behaviour, such as excessive test preparation when attainment scores and public examination results are used to assess schools. Some of these schools may incorporate structural changes but few appear to be making extensive or deep efforts to rethink their instructional programmes or develop capacity.

Feedback during inspection visits

Theories on learning and improvement of schools point to the role of performance feedback in change in schools. During visits, inspectors assess educational quality of schools with respect to standards in a framework and give feedback on the strong and weak points of the performance on these standards. Some Inspectorates also give schools advice on how to improve. Inspection feedback is expected to lead to effects as schools are made aware of the standards they have to comply to and are provided with feedback and sometimes also support to meet these standards.

Inspection visits are also important in preventing strategic behaviour in schools. During visits, inspectors may for example check the accuracy of information presented by the school or may explain to schools how to make genuine improvements to meet the standards. Inspectorates that make little use of inspection visits to assess schools (for example because of long cycles of school inspections or when using other methods to assess schools) are therefore expected to bring about less improvement and cause more

strategic behaviour than Inspectorates visiting schools, particularly when schools have a high stake in being assessed positively.

Ehren, et al. (2013) highlight three hypothetical mechanisms to explain how school inspections lead to school improvement on the classroom and school organisational level:

- Setting expectations and institutionalisation of norms;
- Accepting and using feedback;
- Sensitivity of stakeholders to inspection reports (voice, choice and exit).

Ehren et al. (2013) describe how inspection standards are assumed to *set expectations* about good education for schools and their stakeholders. These authors describe how standards are set when schools take heed of the information included in inspection standards and procedures, reflect on it, process it and adapt their goals and their practices in such a way that they come closer to the desired image of schools communicated by the inspection. Similar processes are described by Segerholm (2011: 1) who also explains how:

evaluative activities, and perhaps specifically if they are carried out systematically, regularly and comprehensively like school inspections, impact on our perception and understanding of ourselves and the surrounding world in particular ways that are expressed in the values permeating these activities.

Standards are set when schools pay attention to the information included in inspection standards and procedures, reflect on it, process it and adapt their goals and their practical ways of working in such a way that they come closer to the standard image of schools portrayed by the inspection framework.

Feedback from school inspections is situated at the interpersonal level (school inspector to principal and/or teacher), as well as within a more complex multi-level system of feedback in inspection reports to the school's stakeholders and publication of summary inspection assessments to the wider public. Inspection feedback becomes relevant when it supports actors on various levels in providing ideas for improvement actions. A number of studies note that *feedback* following an inspection has a great impact on school improvement, particularly when the feedback is specific, frequent and adapted to the school context (Matthews and Sammons, 2005; Ehren and Visscher, 2008; McCrone et al., 2007, 2009; Nusche et al., 2011; Dederling and Muller, 2011; Dobbelaer et al, 2013). Brimblecombe, Shaw, and Ormston (1995) and Chapman (2001), for example, describe how teachers seem to regard oral and written feedback from school inspectors as an important stimulus for school improvement. They found that teachers value the feedback from school inspectors as an important impetus for school improvement activities, especially if given in a setting of trust instead of in a context of punishment. Standaert (2000) confirms these findings when describing how feedback given in a private setting and fitting with a school's culture seems to have a particularly positive impact. Ouston, Fidler, and Earley (1997) point out that school inspections promote greater school improvement if the inspection report clearly details the areas in which the school has performed poorly. According to Matthews and Sammons (2004), clear and explicit reports and feedback to schools are effective in informing school improvement plans after school inspections. Ehren and Visscher (2008) emphasize that feedback in itself often does not lead to improvement, but models of operation where feedback is combined with unsatisfactory scores, specific improvement suggestions and inspection agreements on improvement do make a difference to school improvement.

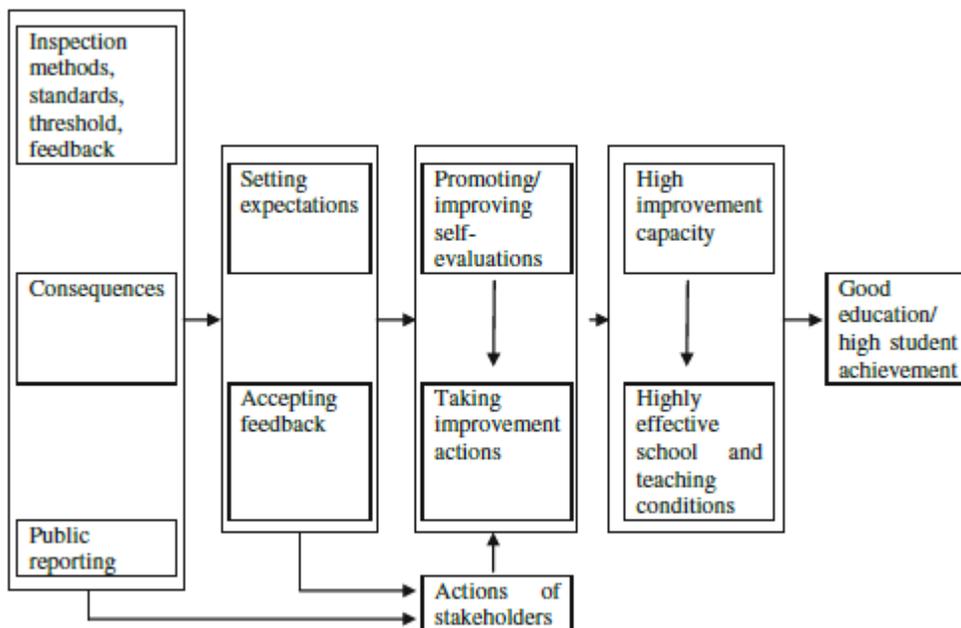
Most inspection models deliberately communicate inspection standards and reports to the school's *stakeholders* (such as parents, local policymakers or school boards). They expect them to use these standards when voicing their opinion about schools, when choosing a school and motivating schools to

address weaknesses and improve. These actions of stakeholders are expected to reinforce inspection expectations and make it more likely that schools react to the inspection. Stakeholders may raise their ‘voice’ in order to motivate schools to improve, or may retreat to the option of ‘choice’ or ‘exit’ where they choose to enter or move their child to a high performing school. ‘Choice’ and ‘exit’ are expected to exert pressure on schools to conform to inspection standards and results by virtue of competition and market pressure.

Each of these three mechanisms operates at multiple levels within the overall system and in the relationship of the system to external stakeholders (e.g., community members, politicians, policymakers). This paper’s interest is in examining the extent to which these mechanisms produce school-level outcomes on the classroom and school organizational level and how these mechanisms change over time to produce positive outcomes.

The following framework summarizes the paper’s assumptions. It outlines how school inspections, their criteria and procedures in general, the consequences of inspection assessments, and the feedback given during inspection visits are expected to enable schools and their stakeholders to align their views/beliefs and expectations of good education and good schools to the standards in the inspection framework, particularly with respect to those standards the school failed to meet during the latest inspection visit. Schools are expected to act on these views and expectations and use the inspection feedback when conducting self-evaluations and when taking improvement actions. Stakeholders should use the inspection standards, or rather the inspection assessment of the school’s functioning against these standards (as publicly reported), to take actions that will motivate the school to adapt their expectations and to improve. Self-evaluations by schools are expected to build their capacity to improve that will lead to more effective teaching and learning conditions. Likewise, improvement actions will (when successfully implemented) lead to more effective school and teaching conditions. These conditions are expected to result in high student achievement.

Figure 1. Conceptual model of intended effects of Dutch school inspections



Methodology

A survey was completed by principals and teachers in Dutch primary and secondary schools in three subsequent years (September – December 2011, 2012 and 2013) to test this model and to identify the mechanisms linking school inspections to the improvement of schools.

Data collection

The online questionnaire to principals and teachers included questions on the intermediate mechanisms of inspection (setting expectations, accepting feedback, stakeholders’ sensitivity to reports) and the intermediate outcome variables (promoting/improving self-evaluations, taking improvement actions, improvement capacity, effective school and teaching conditions) in the theoretical framework. Items to measure the improvement capacity of schools and improvement actions were inspired by the Dutch School Improvement Questionnaire (see Geijsel et al., 2009). Items measuring effective school and teaching conditions were inspired by Scheerens (2009) and were adapted from the ICALT questionnaire which was developed by the Inspectorates of Education in several European countries to measure the quality of teaching and learning, using a shared framework of indicators. Questions on intermediate processes were inspired by the NfER survey ‘Evaluation of the impact of Section 5 inspections’ (McCrone et al, 2009). Principals scored items on the effective school and teaching conditions in their school and the intermediate processes on a 5-point scale ranging from ‘strongly disagree’ to ‘strongly agree’. Questions about improvement actions refer to actions the school has taken to develop its capacity to improve and specifically to enhance effective school and teaching conditions; questions are framed in terms of the amount of time principals have spent during the previous academic year to improve the school’s functioning in these areas (using a 5-point scale ranging from ‘much less time’ to ‘much more time’).

Unintended consequences were measured in the survey via the responses to four items:

Q46: I discourage teachers to experiment with new teaching methods that do not fit the scoring rubric of the Inspectorate

Q47: School inspections have resulted in narrowing curriculum and instructional strategies in my school

Q49: The latest documents/facts and figures we sent to the Inspectorate present a more positive picture of the quality of our school than how we are really doing

Q50: Preparation for school inspection is mainly about putting protocols and procedures in writing that are in place in the school and gathering documents and data.

Principals were asked all four questions in each year of the survey. Teachers were asked all four questions in at least one sweep of the survey, however only Question 46 and Question 47 were repeated in all three sweeps of the teacher survey.

Confirmatory factor analysis was carried out for each latent variable/scale to assess fit. Model fit was deemed to be acceptable or good based on model fit indices (apart from the scale on unintended consequences), see for a full report <authors>. The table below provides a description of the scales.

Table 1. Survey scales and examples of items

Latent construct	Example item	Number of items	Scale
Setting expectations	The inspection standards affect the evaluation and supervision of teachers.	6	strongly agree (1) - strongly disagree (5)
Stakeholders sensitive to reports	The school’s Board of Management/ Boards of Governors is very aware of the contents of the school inspection report.	3	
Accepting feedback	The feedback received from the inspectors was useful.	4	

Promoting/improving self-evaluation	Compared to last academic year, I spent less/more time on the self-evaluation process as a whole.	3	much less (1) - much more (5)
Improvement in capacity building (items on teacher participation in decision making, teacher co-operation, transformational leadership)	Compared to last academic year, I spent less/more time involving teachers in making decisions about using new teaching methods.	8	
Improvement in school effectiveness (items on assessment of teachers/school, opportunity to learn, assessment of students, structured teaching)	Compared to last academic year, I spent less/more time on improving the extent to which teachers make effective use of teaching time within lessons.	10	much less (1) - much more (5)
Capacity building	Teachers collaborate in organizing and improving their teaching	6	strongly agree (1) - strongly disagree (5)
School effectiveness	Students are provided with sufficient instruction time to reach their potential.	5	
Unintended consequences	School inspections have resulted in refocusing curriculum and teaching and learning strategies in my school	5	

Sample

A two stage sampling design was used to select primary and secondary schools and teachers. The sampling design builds from the categories the Inspectorate of Education uses to classify schools and assign them to different inspection treatments (basic, weak, very weak). Schools in these different categories are confronted with different inspection treatments (basic: no visit, weak and very weak: visits and increased monitoring) and they were expected to respond differently to the variables in the survey. The results from the early warning analysis in May 2011 were used to select schools from different inspection categories.

Primary schools

The sample included principals from 408 primary schools, and in each school three teachers from grades 3, 5 and 8. These teachers face different stakes to implement changes in response to school inspections, as particularly students' test scores in grade 8 are part of the inspection measures. Schools in the weak and very weak inspection treatment categories were over sampled to ensure sufficient response rates. Schools that have not been assigned to an inspection treatment or were not included in the early warning analysis due to failures in the information provided to the Inspectorate (595 schools in total) were excluded from the target sample. Tables 1 and 2 in the appendix provide an overview of the target sample and the response rates of each year of data collection of schools and teachers. Response rates are relatively low (particularly in year 1 and 3), but non response of both principals and teachers is similar across the different inspection categories.

Secondary schools

Secondary education commences at the age of 12 and includes different levels of education which are organized in different departments in one school, or in separate school(s) buildings: VMBO, HAVO, or VWO. Children enter a level based on the advice of their primary school and the results of the national standardized end of primary education (grade 8) test. VMBO lasts four years, from the age of twelve to sixteen and combines vocational training with teaching in reading, writing, mathematics, history, arts and sciences. HAVO (five years) and VWO (six years) offer general education leading to higher education. In

this study only the HAVO and VWO departments (548) were included in the selection of secondary schools; 40% of all students nationally are enrolled in HAVO and VWO.

A two stage sampling design was also used to sample school principals and teachers in secondary education, using the results from the early warning analysis from the Inspectorate of Education in May 2011. HAVO and VWO departments that were not included in the early warning analysis of the Inspectorate or had not been assigned to an inspection arrangement were considered out of scope. The target population of secondary schools was therefore set to 454 schools (including both a HAVO and VWO department). The target sample included almost all HAVO and VWO departments in three different inspection treatments to reach sufficient response rates. Due to the limited number of schools in the ‘very weak’ inspection category, all schools in this category were included in the sample.

Teachers from the lower grades (year 1-3 in both HAVO and VWO) and from the final examination grade who teach Dutch language or Geography were included in the sample. Dutch language is considered to be a core and high stakes subject and teachers are required to teach towards nationally defined curriculum standards; such standards have not been set for Geography. The final examination grades are an important inspection measure of student achievement in secondary schools, and these teachers are therefore expected to perceive school inspections as higher stakes compared to their peers in the lower grades.

Tables 1 and 3 in the appendix provide an overview of the target sample and the response rates of schools and teachers. Response rates for secondary schools in year 1 are very low (approximately 5% for both principals and teachers) and even lacking for schools and teachers in the ‘very weak’ inspection category. The results for secondary education should therefore be interpreted with great caution.

Longitudinal sample

We asked the same set of schools to respond to the same survey three times in a row to measure change in these schools as a result of being assigned to the basic, weak or very weak inspection category. There were 285 schools whereby a principal responded at least once to the survey and 317 schools whereby at least one teacher responded. However, there were only 16 (principal) – 18 (teacher) schools that responded in all three years consecutively. For the principal data set there were however responses from 93 schools whereby the principal responded to at least two years of the survey. For the teacher data set there were 105 schools where teachers responded to at least two years of the survey. These patterns of response are highlighted in table 3 below. The column labelled ‘pattern’ refers to the years of data collection for which the school has a response, “1” refers to the school having a response within the year of data collection, and “.” refers to the school missing data for that year. By restricting the sample to the schools who responded to at least two years of the survey, we are able to make better estimates of what the schools’ responses might have been in the one year in which they did not respond. It is important to recognise that it is probable that schools who responded to more than one sweep of the survey, have differential characteristics to those who only responded to only one sweep of the survey. A comparison of the samples on different observable characteristics however indicated that despite the reduction in sample size, the two samples do not differ by great amounts on many of the observable characteristics (see <website>). However, there are more schools from small towns in the longitudinal sample and the schools in the longitudinal sample have fewer children living in poverty.

Table 2. The pattern of missingness within the longitudinal principal and teacher samples

Frequency	Percent	Pattern
Principals		
36	38%	1 1 .
33	36%	. 1 1

16	17%	1 1 1
8	9%	1 . 1
93	100%	
Teachers		
39	37%	. 1 1
35	33%	1 1 .
18	17%	1 1 1
13	12%	1 . 1
105	100%	

Data analysis

Data analyses included three steps:

- 1) Implementing random effect models to analyse changes in each of the variables in the survey according to inspection category of the school. This analysis indicates if ‘basic’ versus ‘weak/very weak’ schools respond differently when faced with different inspection assessments and treatments.
- 2) A longitudinal path model in which we analyze how the variables affect each other over time according to our conceptual framework in figure 1.

Principal and teacher responses are analysed separately as previous analyses suggested that teachers and principals gave systematically different answers to the survey. There are schools in which only teachers have responded and schools in which only the principal has responded. It was not possible to accurately identify teachers over time, therefore as we are interested in changes on the school and teaching level (and not so much in changes of individual teachers), we collapsed teacher responses to provide a single score representing the average teacher response for a school within a particular year. Also the analyses are conducted on the combined responses of primary schools and secondary schools, as the low response rates in secondary schools don’t allow us to measure change in secondary schools separately. Both types of schools are also inspected under the same framework and inspection methods, which suggests that the assumptions as outlined in our theoretical framework in figure 1 are the same for both types of schools.

The analysis was conducted using Stata version 13 (StataCorp, 2013). Scale scores were constructed by performing factor analysis on the polychoric correlation matrices using the pairwise option. Scale scores were predicted using the regression method (DiStefano, Zhu and Mindrila, 2009; Thurstone, 1934).

Step 1

The analysis of the longitudinal data began with graphical representations of the descriptive statistics on the observed primary data. The overall changes over time were considered and these changes were separated by inspection category (very weak/weak/basic).

Random effect models (also known as multilevel models) with GLS estimation were used to test changes over time and the interaction between inspection category and time. This approach provided the most flexibility with the small sample size and the small number of time points. Furthermore random effect models can alleviate problems with missing data through the use of maximum likelihood methods (Quene and van den Bergh, 2004). The random effect model takes into account the dependence of the observations. Another way to think about this is that time points are nested within schools. The analysis is conducted on the predicted scale scores.

Due to the very small number of schools categorised as ‘very weak’ this category has been combined with the ‘weak’ category to create a binary variable of ‘inspection category’. A main effect of inspection category is included to test whether the initial values of the scales, the intercepts, are influenced by the

inspection category of the school. The interaction between time and inspection category is also included to test whether scores in the scales changed differentially for principals and teachers in schools in different inspection categories.

Step 2

Based on the conceptual framework presented in figure 1, a path model was estimated using the longitudinal data. It was expected that the scales ‘accepting feedback’, ‘setting expectations’ and ‘stakeholder sensitivity’ in year one of the survey would go on to influence ‘improvement actions’ of the school including ‘promoting self-evaluations’, ‘improvements in school effectiveness’ and ‘improvements in capacity building’ in year 2. These improvement actions would then influence the scales of ‘capacity building’ and ‘school effectiveness’ in year 3. We fitted this model on the principal and teacher data separately, and calculated models with and without controls for the inspection category.

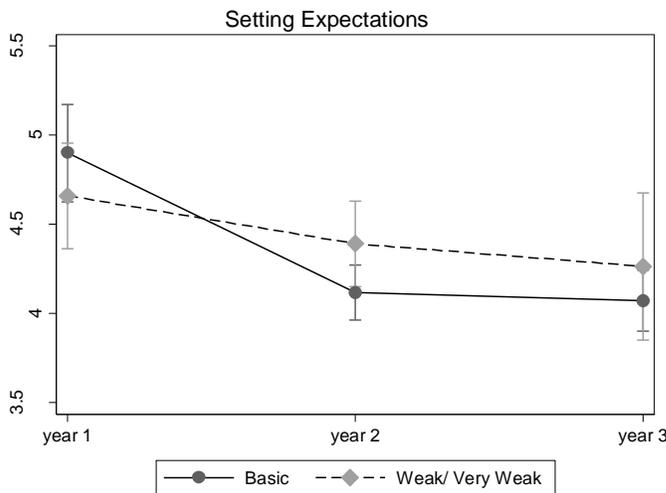
Results: testing changes over time for principals and teachers by inspection category

This section describes the impact of the inspection category upon the responses of the principals and teachers to the scales within the survey over time.

Setting expectations

The results for the setting expectations scale are shown in figure 2. The results indicate that in between year 1 and year 2 of the survey, principals in the “basic” inspection category (year 1 estimate = 4.90, S.E=0.14: year 2 estimate =4.12, S.E=0.08) experienced significantly larger decreases in the setting expectations scale between year 1 and 2 than those in the 'weak/very weak' inspection category (year 1 estimate = 4.65, S.E=0.15: year 2 estimate =4.39, S.E=0.12). However between year 2 and year 3 the different inspection categories follow a similar trajectory, as shown by the similarity of the slopes between year 2 and 3 in figure 2. These findings suggest that the categorization of schools in a basic or ‘weak/very weak’ category leads to a stronger alignment of school and teaching processes to inspection standards in the year after the early warning analysis, but alignment seems to decline over time in all schools. In year 2, principals in schools in the ‘basic’ category score significantly lower on the ‘setting of expectations’ compared to schools in the ‘weak/very weak’ category. These results are somewhat different for teachers: teachers in schools that were in the 'weak/very weak' inspection category tended to report higher scores on the setting expectations scale on average in year 1 (instead of principals in year 2).

Figure 2. Changes in setting expectations for principals in basic versus weak/very weak schools

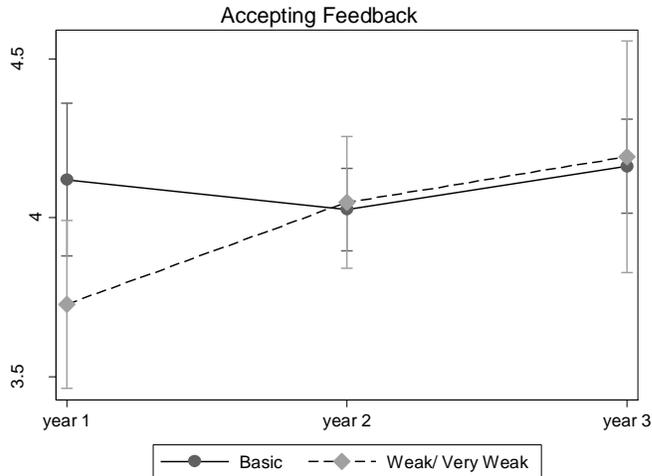


Accepting feedback

The results for the ‘accepting feedback’ scale are shown in figure 3. In the first year of the survey, principals in the ‘weak/very weak’ inspection category (estimate =3.73, S.E=0.14) reported significantly lower scores on average for the ‘accepting feedback’ scale than principals in schools in the basic inspection category (estimate = 4.12, S.E= 0.12). There was also some evidence that between year 1 and year 2 there were differential changes in the ‘accepting feedback’ scores by inspection category: principals in the basic category tended to report slightly lower scores at year 2 compared to year 1 (year 2 estimate =4.03, S.E=0.07), whereas in the ‘weak/very weak’ inspection category the average score increased (year 2 estimate = 4.05, S.E=0.11). Between year 2 and year 3 there are no differences between the groups, which is clearly depicted in figure 3. These findings suggest that, after an inspection visit, principals in the ‘weak/very weak’ inspection category initially do not accept the inspection feedback, but increasingly do so in the second year after the inspection visit. There is however no evidence that the

average teacher response to the accepting feedback scale varied over time by inspection category of the school.

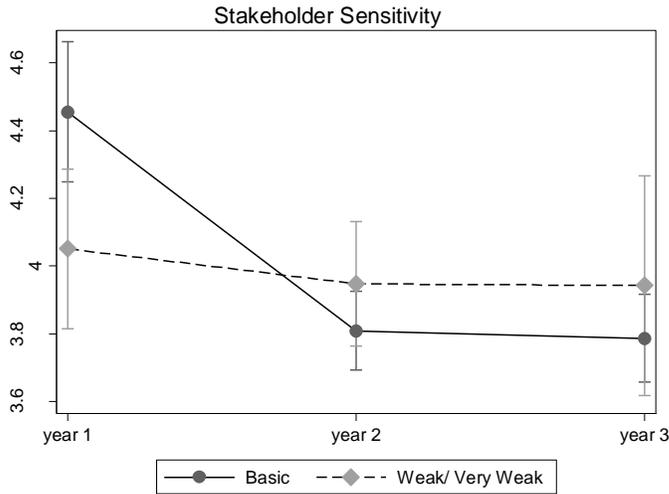
Figure 3. Changes in accepting feedback for principals in basic versus weak/very weak schools



Actions of stakeholders

The results for ‘stakeholder sensitivity’ are shown in figure 4. There was a significant interaction between changes in the ‘stakeholder sensitivity’ scale over time and the inspection category of the school ($\chi^2(2)=8.52, p<0.05$). Principals of schools in the ‘weak/very weak’ inspection category tended to report very similar scores on the ‘stakeholder sensitivity’ scale over time (year 1 estimate=4.05, S.E=0.12; year 2 estimate=3.95, S.E=0.09; year 3 estimate=3.94, S.E=0.17), as shown by the approximation of a straight line in figure 4. Whereas, on average, principals of schools in the ‘basic’ category reported a reduction in scores on this scale over time, particularly between year 1 and 2 (year 1 estimate=4.45, S.E=0.11; year 2 estimate=3.81, S.E=0.06; year 3 estimate= 3.79, S.E=0.07). In year 1 the basic inspection group had significantly higher scores on average than the ‘weak/very weak’ inspection group. But, in year 2 and year 3 they tended to report lower scores on average than the ‘weak/very weak’ inspection group. These findings suggest that particularly stakeholders, such as parents and school boards in potentially high performing schools use the inspection assessment from the early warning analysis in or after the year of publication, but not anymore in subsequent years. There is however no evidence that the average teacher response to the stakeholder sensitivity scale varied over time by inspection category of the school.

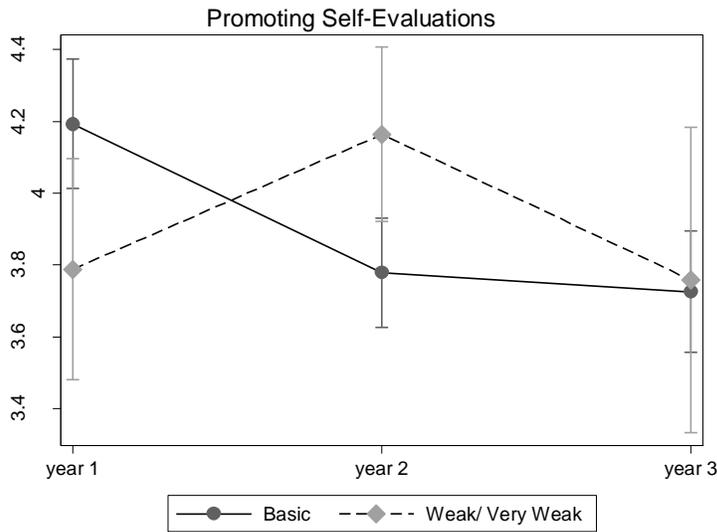
Figure 4. Changes in stakeholder sensitivity for principals in basic versus weak/very weak schools



Promoting self-evaluations

The results for the ‘promoting self-evaluations’ scale are shown in figure 5. There was a significant interaction between changes in the ‘promoting self-evaluations’ scale over time and inspection category ($\chi^2(2) = 12.35, p < 0.05$). Principals in schools in the inspection category ‘basic’, tend to report lower scores for ‘improvement of self-evaluations’ between year 1 and year 2, and increasing scores between year 2 and year 3 (year 1 estimate=4.19, S.E=0.09; year 2 estimate=3.78, S.E=0.08; year 3 estimate=3.73, S.E=0.09). However the exact opposite pattern occurs for principals in schools in inspection category ‘weak/very weak’ (year 1 estimate=3.79, S.E=0.16; year 2 estimate=4.16, S.E=0.12; year 3 estimate=3.76, S.E=0.22). In year 1 and 2 the inspection categories have statistically significant differences in their responses to the ‘promoting self-evaluations’ scale, but by year 3 there are no longer differences by inspection category. These findings suggest that inspection visits in weak and very weak schools lead to additional actions of principals to improve the school’s self-evaluation, particularly in the first two years after the early warning analysis, while potentially high performing schools in the ‘basic’ inspection category have implemented/worked on their self-evaluations before or during the early warning analysis of the Inspectorate, and decrease their level of activity when the Inspectorate publishes the outcome of the early warning analysis and places them in the ‘basic’ category. We did not find any evidence that the average teacher response to the ‘improvement of self-evaluations’ scale varied over time by inspection category of the school.

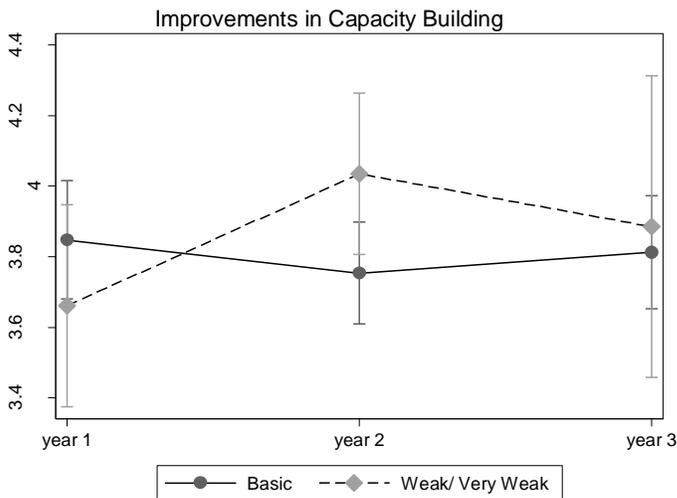
Figure 5. Changes in promoting self-evaluations for principals in basic versus weak/very weak schools



Improvements in capacity-building

The results for the ‘improvements in capacity building’ scale are shown in figure 6. There is a borderline significant association between changes in principals’ scores on the ‘improvement in capacity building’ scale over time and the inspection category of the school ($\chi^2(2)=5.02, p=0.081$). At year 2 there is a significant difference in the scores, with principals in schools in inspection category ‘weak/very weak’ reporting higher scores on average than those in the ‘basic’ inspection category. There are no statistically significant differences at year 1 or year 3 of the survey. These findings suggest that principals particularly invest in cooperation between teachers, transformational leadership, or teachers’ participation in decision-making in the year after an inspection visit. There is no evidence that the average teacher response to the improvements in capacity building scale varied over time by inspection category of the school.

Figure 6. Changes in improvements in capacity-building for principals in basic/weak/very weak schools



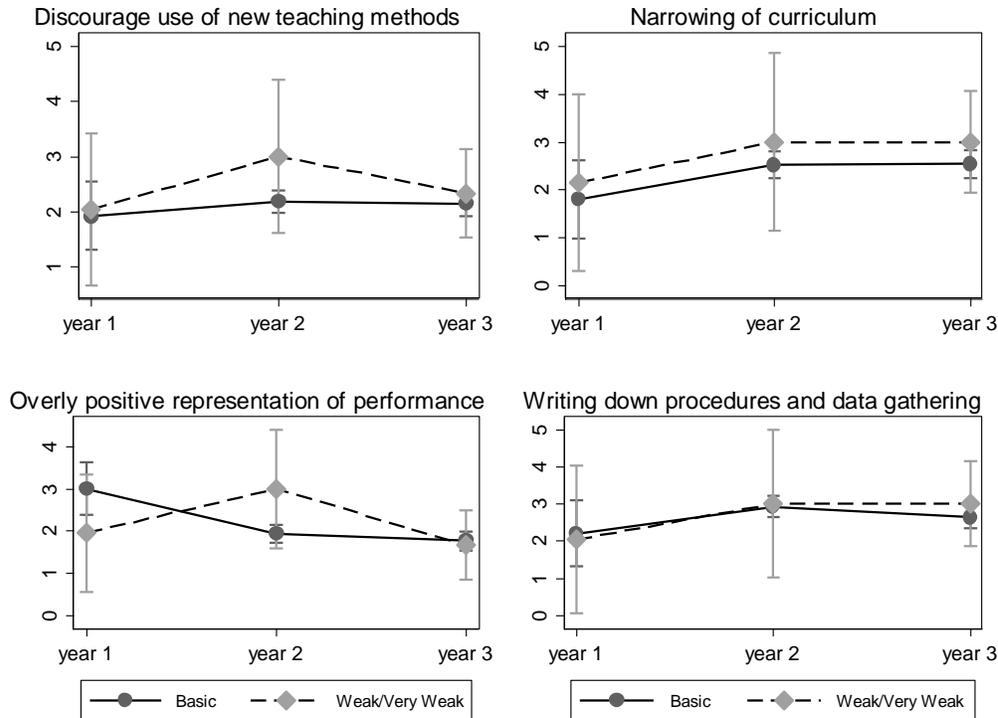
School-effectiveness and capacity-building

There is no evidence that principals responded differently over time to changes in the ‘improvements in school effectiveness’ or the ‘school effectiveness’ scale or ‘capacity-building’ scale, according to the inspection category of their school. Also, teachers’ responses did not vary over time by inspection category of the school on the ‘school effectiveness’ scale. Teachers in schools in the inspection category ‘weak/very weak’ however reported on average higher scores on the ‘improvements in school effectiveness’ scale in year 1, compared to teachers in schools in the ‘basic’ inspection category. Also, teachers in schools in inspection category ‘weak/very weak’ tended to report decreases in ‘capacity building’ over time whereas teachers in schools in the ‘basic’ inspection category reported very similar scores to the ‘capacity building’ scale over time.

Unintended consequences

For principals there were no significant differences in responses to the questions on discouraging the use of new teaching methods, whether inspections lead to a narrowing of the curriculum, and whether school inspection are mainly about putting protocols and procedures in writing that are in place in the school and gathering documents and data. There was also no evidence that principals in schools in different inspection categories responded differently to these items over time. With regard to the question on whether the documents being sent to the inspectorate presented a more positive picture of the school than was true, there was some evidence that principals in schools in different inspection categories responded differently at year 2 of the survey, with principals in schools in the basic inspection category more likely to disagree with the item (basic year 2 estimate = 1.93, S.E=0.10; weak/very weak year 2 estimate=3.00, S.E=0.72).

Figure 7. Changes in unintended consequences for principals in basic/weak/very weak schools



There is a significant interaction between year of survey and inspection category for the average teachers' responses to whether inspections result in a narrowing of the curriculum ($\chi^2(2)=6.81$, $p<0.05$). Responses of teachers in schools in the basic inspection category change very little over the three years of the survey (year 1 estimate=2.75, S.E=0.13; year 2 estimate=2.81, S.E=0.10; year 3 estimate = 2.89, S.E=0.11). Teachers in schools in the weak inspection category are more likely to disagree that inspections result in a narrowing of the curriculum in year 2 and 3 compared to teachers in schools in the basic inspection category (year 1 estimate=3.24, S.E=0.26; year 2 estimate=2.48, S.E=0.19; year 3 estimate =2.35, S.E=0.33). There was however no evidence that teachers in schools in different inspection categories responded differently to the question of whether they were discouraged from using new teaching methods.

Results: Longitudinal Path Models

The section above provides an indication of when changes take place in relation to the timing of the placement of schools in the 'basic' versus 'weak/very weak' inspection category and how these changes are particularly reported by principals. In this section we will use the responses from principals to look at the change process over time, looking at how variables affect each other year on year. We fitted path models, taking into account our assumptions of temporal ordering in the relationship between variables to look at how variables are related and impact on one another, based on the assumptions of our conceptual framework. Cross sectional structural equation models have been fitted using data from the first and second years of the survey looking at these relationships at a single point in time (<authors>). Here, we extend these models to include the inspection categories and the presumed temporal ordering of the relationships. However the smaller sample size and the complexity of the model limits us to using path models rather than structural equation models. We will test if and how 'setting expectations', 'accepting feedback', 'stakeholder sensitivity' impact on 'improving self-evaluations' and 'improving capacity-building', and how these two variables impact on the school's capacity to improve and the effective school and teaching conditions.

The results from the path model are shown in figure 7. The coefficients shown on the regression pathways are standardised coefficients which are standardised on the dependent and independent variables. Being categorised as 'weak/very weak' in the first year of the survey resulted in significantly higher scores (approximately 1/3 of a standard deviation) for improvements in capacity building in year 2. Being categorised as 'weak/very weak' in the second year of the survey was associated with higher scores (approximately 1/3 of a standard deviation) in the 'capacity building' scale. Being categorized as 'weak/very weak' was also indirectly positively associated with 'improvements in school effectiveness' in year two through 'improvements in self-evaluations'. Higher scores on 'promoting self-evaluations' are associated with higher scores on 'improvements in school effectiveness', a standard deviation increase on the 'improving self-evaluations' scale results in just over half a standard deviation increase in the 'improvements in school effectiveness' scale. Also a standard deviation increase in the 'improving self-evaluations' scale is associated with 0.4 of a standard deviation increase on the 'improvements in capacity building' scale. Higher scores in 'improvements in capacity building' in year 2 were associated with lower scores in 'school effectiveness' in year 3. 'Capacity building' is positively associated with 'school effectiveness' at year 3 with higher scores on reported 'capacity building' associated with higher scores on 'school effectiveness'. There is an indirect negative association between 'promoting self-evaluations' in year 2 and 'school effectiveness' in year 3. The full list of coefficients are included in table 4-6 in the appendix.

Figure 8. Longitudinal path model principals. Standardised (X and Y standardisation) coefficients shown

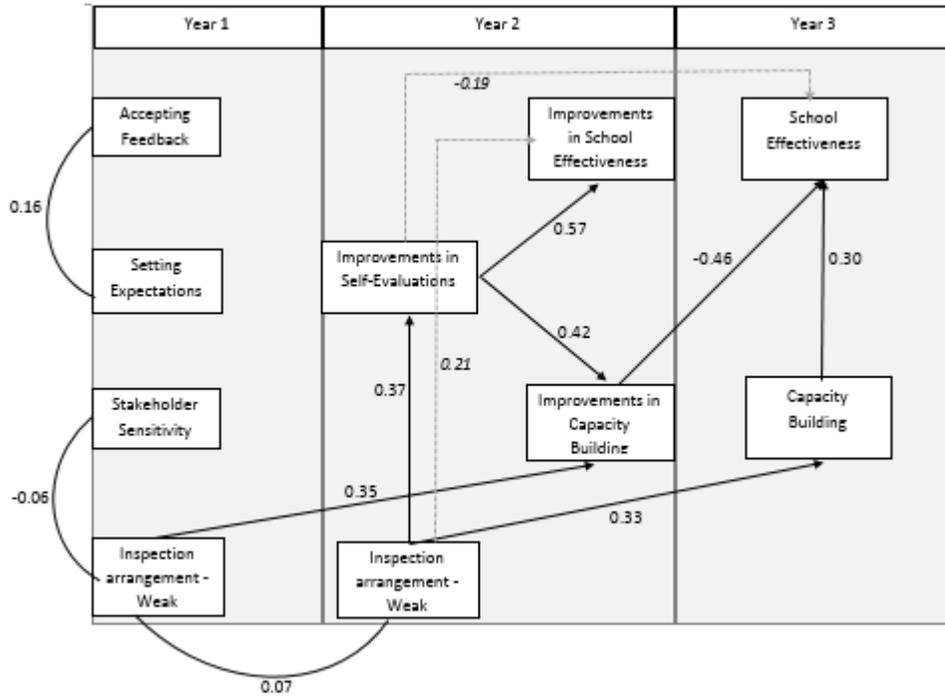


Figure notes; significant indirect effects shown with dashed line.

Conclusion and discussion

This paper used a longitudinal survey to principals and teachers in primary and secondary schools to analyse how school inspections lead to change in schools. The results indicate how schools that receive different inspection treatments change differently over time, and how different mechanisms of change lead to improvement of school and teaching conditions. The analyses made use of a theoretical framework outlined by Ehren et al (2013), describing how school inspections, their criteria and procedures in general, the consequences of inspection assessments, and the feedback given during inspection visits enable schools and their stakeholders to align their views/beliefs and expectations of good education to the standards in the inspection framework, particularly with respect to those standards they failed to meet during the latest inspection visit. According to this model, schools are expected to act on these expectations and use the inspection feedback when conducting self-evaluations and when taking improvement actions. Stakeholders are also expected to use the inspection standards and report of the school's functioning against these standards to take actions that will motivate the school to improve. Self-evaluations by schools are expected to build their capacity to improve that will lead to more effective teaching and learning conditions. Similarly, improvement actions will (when successfully implemented) lead to more effective school and teaching conditions. These conditions are expected to result in high student achievement.

This study is unique in using a longitudinal approach to study these changes from school inspections over time. Implementing random effects models and a longitudinal path model for principals and teachers separately allowed us to analyze how variables impact on each other on different (school organisation and teaching) levels over time. This approach is unique as it allows, for the first time in the study of school inspections, to look at mechanisms of impact and how change from school inspections comes about.

The study however also has some limitations that need to be addressed, particularly in using self-reports of principals and teachers in analysing change, and in the low response rates on the survey and the limited overlap in responses over the three years. We tried to address the issue of self-reports by asking principals and teachers factual questions about the type and level of change they implemented over the last year. The low response rates were taken into account when defining the longitudinal sample for analysis and choosing random effects models with maximum likelihood estimation to estimate missing responses.

These random effect models allowed us to look at changes over time as reported by principals and teachers in schools in the 'basic' and 'weak/very weak' inspection categories. The results suggest that school inspections particularly have an impact on principals, and not so much teachers. Although due to the necessity to average teacher responses within each time point for each school, it was always going to be more difficult for differences to reach significance in the teacher sample. Principals in all schools indicate that school inspections set expectations in the first year after the early warning analysis, but declining scores for all schools suggest that this effect levels off in later years. Schools in the 'basic' inspection category accept significantly more feedback in the first year after the early warning analyses, compared to schools in the 'weak/very weak' category; acceptance of feedback however increases over the three years for schools in the 'weak/very weak' category while it remains relatively stable for schools in the 'basic' inspection category. There is a strong correlation between acceptance of feedback and setting of expectations in the first year after the early warning analysis, suggesting that schools that incorporate inspection standards also accept more inspection feedback.

It was also found that stakeholders from schools in the 'basic' category were sensitive to inspection reports in the first year after the categorization of the school, while they do not seem to use reports anymore in subsequent years. Schools in the 'weak/very weak' inspection category do not report changes in stakeholders' sensitivity to inspection reports.

Schools in the two categories did not report differences in the level of improvements of effective school and teaching conditions or differences in the level of improvement capacity and school effectiveness. We

also found very little differences between schools in the ‘basic’ and ‘weak/very weak’ inspection categories in unintended consequences of inspections, although principals in ‘weak/very weak schools’ more often report of sending documents that represent a more positive picture of the school to the Inspectorate. Teachers in ‘weak/very weak’ schools however also report of less narrowing of the curriculum compared to teachers in schools in the basic inspection category.

Schools in the ‘basic’ category show a decline in their improvement of self-evaluations, while ‘weak/very weak’ schools increase their efforts in the first two years after the early warning analysis, but return back to normal in year 3. Schools in the ‘weak/very weak’ category also improve their capacity in the first two years after the categorization of the school, while schools in the ‘basic’ category remain relatively stable over time.

The results from the path model indicate that improvements in self-evaluations lead to increased activity in improvement of the effectiveness of the school and in improving the school’s capacity in the following year. These improvement actions however did not lead (according to principals) to a more effective school or to more capacity to improve in the school.

These results lead to four main conclusions:

- First, the results indicate different mechanisms of potential impact for schools in different inspection categories: potential improvement from school inspections in the ‘basic’ inspection category seems to result from the setting of expectations and the preparation and improvement of self-evaluations, openness to inspection feedback and sensitivity of stakeholders to inspection reports in the year of, and after the early warning analysis. Weak and very weak schools show a pattern of impact through an increase in openness to, and acceptance of inspection feedback and increasing changes in the schools’ self-evaluations and capacity-building over the years.
- Second, school inspections seem to primarily have an impact on principals and not so much on teachers.
- Third, the results indicate that the actual impact on improved school and teaching conditions is limited. However, as such effects are more likely to take effect after a longer period of time than the three years of data collection, these improvements may potentially be out of the scope of the study.
- Fourth, we find little unintended consequences from school inspections. Only principals in ‘weak/very weak schools’ report of sending documents that present a more positive picture of the school to principals.
- Finally, the lack of any correlation between accepting feedback, setting expectations and stakeholder sensitivity on the one hand and improvement actions in the schools on the other hand also suggests that impact of school inspections is not a linear process, but operates through diffuse and cyclical processes of change.

This paper is an initial attempt to highlight the types of processes that underlie such complex and interrelated mechanisms of impact of school inspections, suggesting that school inspection models can be improved when policy-makers develop models that impact on the teaching level (e.g. through an evaluation of, and feedback about the quality of teaching), and thinking of broader mechanisms of impact through norm-setting and dissemination of good practices.

Evidence from school effectiveness studies supports such a focus and provides further suggestions on the conditions of impact that Inspectorates of Education should focus on, as well as the research models to use in analysing such an impact. Creemers and Kyriakides’ (2008) and Sammons et al (2011) dynamic model of educational effectiveness for example shows how effective schooling is a dynamic process where conditions of school effectiveness operate on different levels (student, class, school organization and school context) and where the relations between those levels and the interplay between the conditions of effective schools on those levels is often curvilinear; e.g. a minimal level of knowledge is necessary for teachers to be effective, but beyond a certain point, a negative relation occurs. A similar approach is

needed for the study of the impact of school inspections, using a dynamic model to analyse the improvement from inspections and for example look at the trajectories of change in schools in different inspection categories. Equally, Inspectorates of Education should allow their inspection frameworks to be more flexible and adaptive to the context in which schools function and their trajectory of improvement.

References

Allen, R. & Burgess, S. (2012). How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England. Bristol, University of Bristol, Centre for Market and Public Organisation, Bristol Institute of Public Affairs.

<Authors>

Brimblecombe, N., Ormston, M., & Shaw, M. (1995). Teachers' Perceptions of School Inspection: a stressful experience. *Cambridge Journal of Education*, 25(1), 53-61. doi: 10.1080/0305764950250106

Chapman, C. (2001). Changing Classrooms through Inspection. *School Leadership & Management*, 21(1), 59-73.

Creemers, B., & Kyriakides, L. (2010). School factors explaining achievement on cognitive and affective outcomes: Establishing a dynamic model of educational effectiveness. *Scandinavian Journal of Educational Research*, 54(3), 263-294.

De Grauwe, A. (2007). Module 1; *Supervision, a key component in a quality monitoring system*. http://www.iiep.unesco.org/fileadmin/user_upload/Cap_Dev_Training/Training_Materials/Supervision/SUP_Mod1.pdf

Dederig, K., & Muller, S. (2011). School Improvement through Inspections? First Empirical Insights from Germany. *Journal of Educational Change*, 12(3), 301-322.

de Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of Education*, 33(3), 379-396. doi: 10.1080/03054980701366207

DiStefano, C., Zhu, M. and Mindrila, D. (2009). 'Understanding and using factor scores: Considerations for the applied researcher'. *Practical Assessment, Research & Evaluation*, 14 (20), 1-11.

Dobbelaer, M. J., Prins, F. J., & van Dongen, D. (2013). The Impact of Feedback Training for Inspectors. *European Journal of Training and Development*, 37(1), 86-104.

Ehren, C. M., & Visscher, J. A. (2008). The Relationships between School Inspections, School Characteristics and School Improvement. *British Journal of Educational Studies*, 56(2), 205-227.

Ehren, C. M., Altrichter, H., McNamara, G., & O'Hara, J. (2013). Impact of School Inspections on Improvement of Schools--Describing Assumptions on Causal Mechanisms in Six European Countries. *Educational Assessment, Evaluation and Accountability*, 25(1), 3-43.

Elmore, R.F. and Fuhrman, S.H. (2001). Research Finds the False Assumption of Accountability. *Phi Delta Kappan*, 67(4), 9-14.

Geijsel, P., Slegers, P.J.C., Stoel, R.D. and Kruger, M.L. (2009). The Effect of Teacher Psychological and School Organizational and Leadership Factors on Teachers' Professional Learning in Dutch Schools. *The Elementary School Journal*, 109(4), 1-22.

- Gray, A. (2014). Supporting school improvement: the role of inspectorates across Europe. *Brussels: SICI*. <http://www.sici-inspectorates.eu/getattachment/5caebee9-84c1-41f0-958c-b3d29dbaa9ef> (retrieved July 2014)
- Grek, S., Lawn, M., Ozga, J., & Segerholm, C. (2013). Governing by inspection? European inspectorates and the creation of a European education policy space. *Comparative Education*, 49(4), 486-502.
- Hanushek, E.A. and Raymond, M.E. (2002). Lessons about the Design of State Accountability Systems. *Paper prepared for 'Taking Account of Accountability: Assessing Policy and Politics'*, Harvard University.
- Heubert, J.P. and Hauser, R.M. (Eds.) (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington: National Academy Press.
- Hussain, I. (2012). *Subjective performance in the public sector: evidence from school inspections*: London School of Economics and Political Science. Centre for Economic Performance.
- Kerr, S. (1975). On the folly of rewarding A, while hopping for B. *The Academy of Management Journal*, 18(4), 769-783.
- Klerks, M. (2013). The effect of school inspections: a systematic review. www.schoolinspections.eu (retrieved January 2014).
- Koretz, D.M. (2003). Using Multiple Measures to Address Perverse Incentives and Score Inflation. *Educational Measurement*, 22(2), 18-26.
- Looney, J. (2009). *Assessment and innovation in education*. OECD Education Working Paper No. 24 (EDU/WKP(2009)3).
- Luginbuhl, R., Webbink, D., & de Wolf, I. (2009). Do Inspections Improve Primary School Performance? *Educational Evaluation and Policy Analysis*, 31(3), 221-237.
- Malen, B. (1999). On Rewards, Punishments, and Possibilities: Teacher Compensation as an Instrument for Education Reform. *Journal of Personnel Evaluation in Education*, 12(4), 387-394.
- Matthews, P. & Sammons, P. (2004). *Improvement through Inspection: An Evaluation of the Impact of Ofsted's Work*. London: IOE/Ofsted, Document reference number: HMI 2244
- Matthews, P., & Sammons, P. (2005). Survival of the Weakest: The Differential Improvement of Schools Causing Concern in England. *London Review of Education*, 3(2), 159-176.
- McCrone, T., Coghlan, M., Wade, P., & Rudd, P. (2009). Evaluation of the impact of Section 5 inspections - strand 3. Final report for Ofsted. *Evaluation of the impact of Section 5 inspections - strand 3. Final report for Ofsted*, 58.
- Nelson, R. and Ehren, M.C.M. (2014). Review and synthesis of evidence on the (mechanisms of) impact of school inspections. <http://schoolinspections.eu/wp-content/uploads/downloads/2014/02/Review-and-synthesis-of-evidence-on-the-mechanisms-of-impact-of-school-inspections.pdf>

Nichols, S.L. and Glass, G.V. and Berliner, D.C. (2006). High-stakes testing and student achievement: does accountability pressure increase student learning? *Education Policy Analysis Archives*, 14(1), Retrieved 14 November 2008 from <http://epaa.asu.edu/epaa/v14n1>.

Nusche, D., Halász, G. G., Looney, J., Santiago, P., & Shewbridge, C. (2011). OECD Reviews of Evaluation and Assessment in Education: Sweden 2011.

OECD (2013), *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*, OECD Reviews of Evaluation and Assessment in Education, OECD Publishing. doi: [10.1787/9789264190658-en](https://doi.org/10.1787/9789264190658-en)

Ouston, J., Fidler, B. & Earley, P. (1997). What do schools do after OFSTED school inspections-or before? *School Leadership & Management*, 17(1), 95-104.

Quene, H. and van den Bergh, H. (2004). 'On multi-level modeling of data from repeated measures designs: a tutorial'. *Speech Communication*, 43 (1-2), 103-121.

Rosenthal, L. (2004). Do School Inspections Improve School Quality? Ofsted Inspections and School Examination Results in the UK. *Economics of Education Review*, 23(2), 143-151.

Sammons, P., Gu, Q., Day, C., Ko, J. (2011). Exploring the impact of school leadership on pupil outcomes: Results from a study of academically improved and effective schools in England. *International Journal of Educational Management*, Vol. 25 Iss: 1 pp. 83 - 101

Segerholm, C. (2011). *Values in Evaluation: the what and how values in Swedish school inspection*. Paper presented at the American Evaluation Association Conference, Anaheim, 2-5 November 2011.

Shaw, I., Newton, P. D., Aitkin, M., & Darnell, R. (2003). Do OFSTED Inspections of Secondary Schools Make a Difference to GCSE Results? *British Educational Research Journal*, 29(1), 63.

Scheerens, J. (2009). Review and Meta-analyses of School and Teaching Effectiveness. The Netherlands/department of Educational Organization and Management. www.iqb.hu-berlin.de/lehre/dateien/rapportScherens.pdf

Standaert, R. (2000). *Inspectorates of education in Europe; a critical analysis*. Flanders: ministry of education.

StataCorp. (2013). Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

Stecher, B.M. (2001). Consequences of large-scale, high-stakes testing on school and classroom practices). Tests and their use in test-based accountability systems. In Hamilton, L.S., Stecher, B.M., Klein, S.P. (Eds.). *Making sense of Test-based Accountability in Education*. Santa Monica: Rand cooperation. http://www.rand.org/pubs/monograph_reports/MR1554/

Thurstone, L. L. (1934). 'The Vectors of Mind'. *Psychological Review*, 41, 1-32.

Wößmann, L., Lüdemann, E., Schütz, G. And West, M.R. (2007). *School Accountability, Autonomy, Choice, and the Level of Student Achievement: International Evidence from PISA 2003*. OECD: EDU/WKP(2007)8.

Table 1 Target sample and response rates, principals and teachers combined

		Year 1		Year 2		Year 3		All 3 years	
Target Sample		Response rates		Response rates		Response rates		Response rates	
Primary	Secondary	Primary	Secondary	Primary	Secondary	Primary	Secondary	Primary	Secondary
411	359	Number of responses: 213	Number of responses: 100	Number of responses: 339	Number of responses: 251	Number of responses: 199	Number of responses: 181	Number of responses: 751	Number of responses: 532
		Number of schools: 96	Number of schools: 40	Number of schools: 166	Number of schools: 100	Number of schools: 117	Number of schools: 95	Number of schools responding once: 148	Number of schools responding once: 113
								Number of schools responding twice: 79	Number of schools responding twice: 40
								Number of schools responding three times: 24	Number of schools responding three times: 16
								Total number of schools: 251	Total number of schools: 161

Table 2. Target sample and response rates principals and teachers primary education according to inspection category

	Primary education							
		Response rate principals				Response rate teachers		
	Target sample of schools (percentage of target population)	Year 1	Year 2	Year 3	Target sample of teachers (percentage of target population: 1 teacher group 3, 5 and 8)	Year 1	Year 2	Year 3
Schools assigned to basic inspection category	208 (3.10%)	46	77	59	624 (1.86%)	50	66	70
Schools assigned to 'weak schools' inspection category	152 (41.53%)	23	37	9	456 (24.92%)	24	29	5
Schools assigned to 'very weak schools' inspection category	51 (83.61%)	2	9	1	153 (50.16%)	4	8	0
<i>Total</i>	411 (6.19%)	71	123	69	1233 (3.70%)	78	103	75

Table 3. Target sample and response rates principals and teachers secondary education according to inspection category

	Secondary education									
	Target sample of schools (percentage of target population)		Response rate principals			Target sample (percentage of target population: 4 teachers in each department)		Response rate teachers		
	HAVO	VWO	Year 1	Year 2	Year 3	HAVO	VWO	Year 1	Year 2	Year 3
Schools assigned to basic inspection category	321 (77.16%)	262 (73.39%)	12	53	47	1284 (15.43%)	1048 (14.68%)	31	68	51
Schools assigned to 'weak schools' inspection category	33 (100%)	91 (100%)	0	0	3	132 (20%)	364 (20%)	2	3	2
Schools assigned to 'very weak schools' inspection category	5 (100%)	6 (100%)	1	1	0	20 (20%)	24 (20%)	1	1	1
<i>Total</i>	359 (79.10%)	359 (79.10%)	13	54	50	1436 (15.81%)	1436 (51.81%)	34	72	54

Table 4. Full list of coefficients for longitudinal path model; direct effects

Direct Effects	Estimate	S.E	z	p	Standardised coefficient
selfevaly2	<-				
setexpy1	0.08	0.26	0.33	0.75	0.08
insarry1	-0.27	0.31	-0.85	0.39	-0.18
insarry2	0.53	0.23	2.27	0.02	0.37
accfeedy1	-0.01	0.32	-0.03	0.98	-0.01
stakeholy1	-0.03	0.40	-0.08	0.94	-0.02
chschooleffy2	<-				
selfevaly2	0.46	0.08	5.64	0.00	0.57
setexpy1	0.04	0.19	0.19	0.85	0.04
insarry1	0.15	0.22	0.67	0.51	0.13
insarry2	-0.07	0.17	-0.43	0.67	-0.06
accfeedy1	0.09	0.23	0.39	0.70	0.09
stakeholy1	0.03	0.28	0.11	0.91	0.02
changecapbuildy2	<-				
selfevaly2	0.36	0.10	3.81	0.00	0.42
setexpy1	0.06	0.19	0.35	0.73	0.07
insarry1	0.45	0.22	2.00	0.05	0.35
insarry2	-0.06	0.18	-0.37	0.72	-0.05
accfeedy1	0.32	0.23	1.42	0.16	0.29
stakeholy1	0.38	0.28	1.36	0.18	0.27
capbuildy3	<-				
selfevaly2	-0.26	0.17	-1.53	0.13	-0.32
chschooleffy2	0.24	0.21	1.16	0.25	0.24
changecapbuildy2	-0.22	0.17	-1.35	0.18	-0.24
insarry1	-0.09	0.17	-0.54	0.59	-0.08
insarry2	0.38	0.18	2.03	0.04	0.33
schooleffy3	<-				
selfevaly2	0.05	0.18	0.31	0.76	0.06
chschooleffy2	0.17	0.21	0.83	0.41	0.15
changecapbuildy2	-0.49	0.16	-2.96	0.00	-0.46
capbuildy3	0.34	0.13	2.55	0.01	0.30
insarry1	0.24	0.17	1.37	0.17	0.18
insarry2	0.36	0.19	1.87	0.06	0.28

Table 5. Full list of coefficients for longitudinal path model; indirect effects

Indirect effects	Estimate	S.E	z	p	Standardised coefficient
chschooleffy2	<-				
setexpy1	0.04	0.12	0.33	0.74	0.05
insarry1	-0.12	0.15	-0.84	0.40	-0.10
insarry2	0.24	0.12	2.06	0.04	0.21
accfeedy1	0.00	0.15	-0.03	0.98	0.00
stakeholy1	-0.01	0.18	-0.08	0.94	-0.01
changecapbuildy2	<-				
setexpy1	0.03	0.09	0.32	0.75	0.03
insarry1	-0.10	0.12	-0.79	0.43	-0.08
insarry2	0.19	0.11	1.74	0.08	0.16
accfeedy1	0.00	0.12	-0.03	0.98	0.00
stakeholy1	-0.01	0.15	-0.08	0.94	-0.01
capbuildy3	<-				
selfevaly2	0.03	0.02	1.17	0.24	0.03
setexpy1	-0.03	0.07	-0.37	0.71	-0.03
insarry1	0.00	0.12	-0.03	0.98	0.00
insarry2	-0.13	0.09	-1.33	0.19	-0.11
accfeedy1	-0.05	0.10	-0.50	0.62	-0.05
stakeholy1	-0.07	0.12	-0.59	0.55	-0.05
schooleffy3	<-				
selfevaly2	-0.18	0.08	-2.29	0.02	-0.19
chschooleffy2	0.08	0.07	1.16	0.25	0.07
changecapbuildy2	-0.08	0.06	-1.35	0.18	-0.07
setexpy1	-0.04	0.10	-0.38	0.71	-0.04
insarry1	-0.21	0.16	-1.34	0.18	-0.16
insarry2	0.08	0.14	0.58	0.56	0.06
accfeedy1	-0.16	0.13	-1.18	0.24	-0.14
stakeholy1	-0.20	0.16	-1.24	0.22	-0.13

Table 6. Full list of coefficients for longitudinal path model; total effects

Total Effects	Estimate	S.E	z	p	Standardised coefficient
selfevaly2	<-				
setexpy1	0.08	0.26	0.33	0.75	0.08
insarry1	-0.27	0.31	-0.85	0.39	-0.18
insarry2	0.53	0.23	2.27	0.02	0.37
accfeedy1	-0.01	0.32	-0.03	0.98	-0.01
stakeholyl	-0.03	0.40	-0.08	0.94	-0.02
chschooleffy2	<-				
selfevaly2	0.46	0.08	5.64	0.00	0.57
setexpy1	0.07	0.25	0.30	0.77	0.09
insarry1	0.03	0.29	0.09	0.93	0.02
insarry2	0.17	0.21	0.79	0.43	0.15
accfeedy1	0.08	0.30	0.28	0.78	0.08
stakeholyl	0.02	0.38	0.04	0.97	0.01
changecapbuildy2	<-				
selfevaly2	0.36	0.10	3.81	0.00	0.42
setexpy1	0.10	0.21	0.46	0.65	0.10
insarry1	0.35	0.25	1.42	0.16	0.28
insarry2	0.13	0.20	0.65	0.52	0.10
accfeedy1	0.32	0.26	1.23	0.22	0.29
stakeholyl	0.37	0.32	1.16	0.25	0.26
capbuildy3	<-				
selfevaly2	-0.23	0.17	-1.35	0.18	-0.29
chschooleffy2	0.24	0.21	1.16	0.25	0.24
changecapbuildy2	-0.22	0.17	-1.35	0.18	-0.24
setexpy1	-0.03	0.07	-0.37	0.71	-0.03
insarry1	-0.09	0.19	-0.48	0.63	-0.08
insarry2	0.25	0.19	1.35	0.18	0.22
accfeedy1	-0.05	0.10	-0.50	0.62	-0.05
stakeholyl	-0.07	0.12	-0.59	0.55	-0.05
schooleffy3	<-				
selfevaly2	-0.12	0.19	-0.64	0.52	-0.13
chschooleffy2	0.25	0.22	1.16	0.24	0.23
changecapbuildy2	-0.56	0.17	-3.25	0.00	-0.53
capbuildy3	0.34	0.13	2.55	0.01	0.30
setexpy1	-0.04	0.10	-0.38	0.71	-0.04

insarry1	0.03	0.22	0.12	0.90	0.02
insarry2	0.44	0.20	2.17	0.03	0.34
accfeedyl	-0.16	0.13	-1.18	0.24	-0.14
stakeholy1	-0.20	0.16	-1.24	0.22	-0.13