Educ Asse Eval Acc
DOI 10.1007/s11092-016-9242-0

Risk-based school inspections: impact of targeted inspection approaches on Dutch secondary schools

Melanie C. M. Ehren[1] · **Nichola Shackleton**[1]

Received: 21 January 2016 / Accepted: 31 May 2016

Abstract In most countries, publicly funded schools are held accountable to one inspectorate and are judged against agreed national standards. Many inspectorates of education have recently moved towards more proportional risk-based inspection models, targeting high-risk schools for visits, while schools with satisfactory student attainment levels are excluded from inspections. This paper looks into these newer inspection models and aims to enhance our understanding of the potential effectiveness of such targeted models on student attainment and other performance indicators. Random effects models, analyzing changes in schools over time, indicate that targeted inspections particularly have an effect on student attainment in literacy in weak schools, while also impacting on student satisfaction, student numbers and student-staff ratios.

Keywords School inspections · School accountability · Educational effectiveness

1 Introduction

Good education is a major public interest and considered to be an important condition for economic growth and general well-being. Many countries, therefore, pay considerable attention to the quality of their schools and try to improve their level of education. School inspection is used by most European education systems as a major instrument for controlling and promoting the quality of schools. As De Grauwe (2007, p. 6) notes:

> School inspections are external evaluations of schools, undertaken by officials outside of the school with a mandate from a national/local authority. Regular visits of schools are an essential part of school inspections to collect information about the quality of the school, check compliance to legislation and/or evaluate

✉ Melanie C. M. Ehren
m.ehren@ioe.ac.uk

[1] UCL Institute of Education, 20 Bedford Way, London WC1H 0AL, UK

Published online: 08 July 2016

the quality of students' work (e.g. through observations, interviews and document analysis)' (De Grauwe 2007, p. 6).

The Dutch Inspectorate of Education was established in 1801 and has seen many changes over the years, particularly in the last decade (Inspectie van het Onderwijs/ Dutch Inspectorate of Education 2012). The most prominent change included the introduction of risk-based inspections in 2008. Risks refer, according to De Wolf and Honingh (2014), to potential problems and dangerous situations, such as failing teaching quality and lack of good governance in schools. Risks are often based on estimates of the gravity and seriousness of an event and the chances of such an event occurring. Risk-based inspections start with an early warning analysis of all the schools, using available information on possible risks of low educational quality in schools, such as student achievement results on standardized tests and school documents. Schools with no risks are not scheduled for inspection visits, whereas schools that show risks receive additional inspection monitoring and interventions. Risk-based school inspections are expected to increase effectiveness of school inspections by identifying potentially low-performing schools and increasing inspection activities in these schools (and less inspection activities in well-performing schools) (Ehren and Honingh 2011).

Thinking about risks as a starting point for school inspections has become more common over the last years, according to De Wolf and Honingh (2014), and reflects the need for more selective and targeted approaches. They reflect a change in the political and economic situation (particularly the economic recession) and the need for more efficient and effective school inspection models. Other European countries have also developed more targeted approaches over the last years, and the Bratislava memorandum of SICI (2013), the European Association of Inspectorates of Education, even describes the need for inspectorates to be agile and focus on risks and proportionality as one of the most important trends in school inspections across Europe. Examples of such targeted models can be found in England where Ofsted adapts the frequency of inspections to perceived need, such that schools judged satisfactory will be on a 3-year cycle and schools judged as unsatisfactory will be visited more frequently (Allen and Burgess 2012). Similar trends were described for Sweden, Northern Ireland, Estonia and Ireland (Gray 2014).

The introduction of these targeted risk-based inspections has, however, not been without criticism. De Wolf and Honingh (2014), for example, debate whether it is possible to measure risks as they often represent 'yesterday's problems'. A risk-based approach implies a strong focus on performance indicators, limiting the potential value of professional judgement of school inspectors. De Wolf and Honingh (2014) describe that the focus on specific risks as a starting point for inspections may bias judgements of inspectors when they visit schools that have been identified to be potentially failing. Early warning analysis and the assessment of potential failures often become self-fulfilling prophecies when inspectors assess schools along the outcome of the early warning analysis, instead of observing school and teaching quality with an open and objective mind.

Ehren and Honingh (2011) also conclude that the focus on achievement in risk-based inspections does not allow for an actual early warning of risks to prevent failing quality in schools. As these authors indicate, the choice of student achievement results

as the primary risk indicator implies an identification of schools that are already failing. Inadequate teaching and leadership in these schools have led to low student achievement (and not the other way around). Warnings to prevent such failings, however, need to identify schools before their student achievement results are below average and should therefore include other indicators on, for example, internal quality control and self-evaluations of schools to identify potential weaknesses in educational quality, indicators on the quality of educational processes or other causes of failure (such as substantive changes in student or teacher population).

The move towards risk-based inspections, focused on student outcomes, also seems to limit the opportunities of school inspections to motivate school improvement as many schools will not receive feedback on the quality of their teaching and school organization anymore. Ehren and Honingh (2011) and De Wolf and Honingh (2014) suggest that the strong and narrow focus on student outcomes in risk-based inspections may even lead to undesirable gaming and manipulation of the performance indicators used in the early warning analyses.

Evidence of such potential strategic responses is, however, limited as current studies on the impact (both intended and unintended) of school inspections only address school inspections, which include regular cyclical visits to and assessments of all schools. This paper therefore aims to enhance our understanding of the impact of these proportionate risk-based models of school inspections on secondary schools. We will explore if these models lead to improved student outcomes as well as improvement of other performance indicators (e.g. student and parent satisfaction) in Dutch secondary schools (higher general secondary education (HAVO)/pre-university secondary education (VWO) track).

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

2 Secondary education in the Netherlands

By the age of 12, Dutch children go to secondary education. Secondary education in the Netherlands is strongly stratified into pre-vocational education (VMBO), individualized pre-vocational education (VMBO and IVBO, age range 12–16 years), senior general secondary education (HAVO, age range 12–17 years) and pre-university education (VWO, age range 12–18 years). The choice for one of these tracks is made by the pupil and its parents, using the advice from the primary school and the results of a national standardized end of primary education test (the CITO-test).

A school building can offer all or just one of these tracks. This study only includes (departments of) schools offering senior general secondary education and pre-university education (HAVO and VWO), which is offered to approximately 40 % of all the students in secondary education. Each educational track in secondary education is generally led by a department head who is full time responsible for the daily functioning and operation of each track and of the quality of education in each track.

EP-Nuffic (2015) describes the Dutch education system as one where the Ministry of Education, Culture and Science is responsible financial structures, general education policy and admission requirements, and structure and objectives of the education

system. Under the authority of the Ministry of Education, learning objectives are specified at the different stages and different tracks of the education system, but schools are autonomous in deciding on appropriate teaching and learning methods and curriculum design as long as they ensure the incorporation of these learning objectives (Béguin and Ehren 2011).

The school-leaving exams at the end of secondary education assess whether students meet the learning objects and qualify to graduate. These exams are composed of a school-based exam (contributing to 50 % of the final grade) and a national standardized test at the end of the final school year (also 50 % of the final grade).[1] Students need to take both school and national examination in seven subjects to qualify for a pre-university (VWO) diploma and six subjects to qualify for a senior general secondary education (HAVO) diploma (EP-Nuffic 2015). Dutch language is a compulsory subject in the national examination in all types of secondary education. English language and mathematics are also compulsory elements in the national examination in pre-university and senior general secondary education. Students need to pass at least two of the three basic subjects (English, Dutch language and mathematics) to qualify for a diploma.

The elements to be tested in each examination are, according to Béguin and Ehren (2011), specified in the examination syllabus, approved by the Ministry of Education, Culture and Science. The syllabus also specifies the number and length of the tests that make up the national examination. Schools are responsible for setting up the school examination; they develop and submit a school examination syllabus to the inspectorate, showing which elements will be tested, when, how marks are calculated, including the weight allocated to these tests and resit opportunities. The school examination must be completed, and the results were submitted to the inspectorate before the national examinations start.

3 School inspections in the Netherlands

One of the hallmarks of the Dutch Education Inspectorate is the development and use of risk-based inspections in 2008. Risk-based inspections have been common in other sectors as a cost-effective method to select appropriate maintenance and inspections tasks and techniques and to ensure a more proactive regime of inspections (Power 1999). Risk-based inspections of schools were considered appropriate in the Dutch context as the autonomy of schools and developing structures of horizontal accountability required a less burdensome inspection system, but one that would allow the inspectorate to quickly intervene in failing schools (see Ehren and Honingh 2011). The risk-based model allowed the inspectorate to use annual early warning to target potentially failing schools and schedule their inspection capacity for intensive monitoring of these schools.

As the inspectorate of education explains (2013[2]), the early warning analyses are implemented at a fixed moment in the year, after the student achievement data on

[1] http://www.ncee.org/programs-affiliates/center-on-international-education-benchmarking/top-performing-countries/netherlands-overview/netherlands-instructional-systems/

[2] http://www.onderwijsinspectie.nl/binaries/content/assets/Documents+algemeen/2013/231213-publieksversie_waarderingskader_vo_2013.pdf and http://www.onderwijsinspectie.nl/binaries/content/assets/Documents+algemeen/2013/toezichtkader-vo-2013_versie-20131101.pdf (in Dutch)

national examinations are made available by the national testing agency. In these analyses, average student achievement results on the national examinations in all subjects over a period of 3 years (corrected for socio-economic background of students), annual performance data (on staff turnover, number of students and financial viability of the school) and number of complaints of parents or in national and local media are used to identify potentially failing schools. Students' results on national standardized examinations are the primary indicator in the analyses and are 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.

Author (2012, p. 6) further describes and summarizes the methodology of risk-based inspections:

> The early warning analysis is used to classify schools in three categories; schools in the 'green; category are considered to have no risks of failing quality, 'orange' schools have potential risks of failing quality, whereas 'red' schools have high risks of failing quality. The Inspectorate of Education schedules desk research in schools in the orange category. These schools are requested to send in the student achievement results of students in intermediate grades in reading and mathematics. The Inspectorate also analyses additional documents of the school, such as annual reports. In case this desk research shows no risks (the documents are in order and the intermediate results are sufficient and there are no indications of risks); the school gets reassigned to the green category. The school board however receives an informal warning in case the achievement of students in the final grade is below average or is declining.
>
> An interview with the school board is scheduled in case the desk research points to potential risks. Potential risks are 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 is scheduled in case this interview does not provide the Inspectorate of Education with sufficient information or in case the capacity of the school board to address the risks is evaluated as inadequate. During this visit, the inspection framework is used to assess educational quality in the school as sufficient, failing or highly underdeveloped.
>
> The Inspectorate of Education also schedules desk research of schools in the red category, comparable to the desk research of schools in the orange category. School boards of schools in the red category are also scheduled for an interview and schools receive a full inspection visit to evaluate their educational quality. Schools that are evaluated as failing or highly underdeveloped are scheduled for additional inspection activities. The Inspectorate of Education instructs the school board to formulate a plan of approach aimed at improving quality. The Inspectorate tests the plan and lays down performance agreements in an inspection plan. This plan specifies when the quality should be up to par again and what (interim) results the school must attain. It also specifies the indicators the Inspectorate of Education will assess in (interim) inspection visits. The school board must commit to the inspection plan. Failing schools that do not improve within two

years end up in the regime imposed on highly underdeveloped schools. These schools are confronted with additional inspection activities, such as a meeting between the school board and the Inspectorate management or an official warning. If these activities do not yield the result agreed upon, the Inspectorate will report a highly underdeveloped school to the Minister, along with a proposal for instating sanctions. On the basis of this report, the Minister may proceed to impose administrative and/or financial sanctions.

4 Theoretical framework

Earlier reviews of Klerks (2013) and Nelson and Ehren (2014) note that little empirical research has been conducted on the impact of inspection, particularly outside the UK and the Netherlands. Of the available studies, only a limited number of studies look into the impact of school inspection on student achievement. None of these studies, however, address improvement of schools on other performance indicators, such as student and parent satisfaction, student-teacher ratios, number of teaching hours or teachers' sick leave; available studies on student achievement also only look into improved attainment in mathematics and literacy and neglect improvement in other subject areas (e.g. geography, biology).

Notable studies on the impact of school inspections on student achievement include Shaw et al. (2003), Rosenthal (2004), Luginbuhl et al. (2009), Allen and Burgess (2012) and Hussain (2012). Shaw et al. (2003) found that achievement in (far) above- and below-average performing English schools slightly improves after an inspection visit. Inspection did not improve examination achievement in maintained comprehensive schools. Rosenthal (2004) even found a decrease in examination results of students in England in secondary education in the year of the inspection visit. He explains this result by pointing to the extensive preparation of schools for the visit that may take time and energy away from the teaching and learning process. Luginbuhl et al. (2009), on the other hand, found an improvement in test scores of students in primary education, particularly in schools with high proportions of disadvantaged students in arithmetic in the first year following an inspection visit.

More recent reports presented by Allen and Burgess (2012) and Hussain (2012) are based on separate, large, longitudinal datasets in England with a sophisticated process for analysis. Both of these studies claim a link between the findings of an inspection report and student achievement results and suggest that a negative inspection judgement may prompt or accelerate actions to improve student performance, even where no external interventions are made. Hussain also examined improvement in relation to prior attainment, to control for 'gaming' by schools, for example, by failing to enter students less likely to perform well or by targeting borderline students. Such types of strategic behaviour have been described by numerous authors (e.g. Smith 1995; De Wolf and Janssens 2007) and would include schools manipulating performance data send to or collected by the inspectorate (e.g. to be included in the early warning analyses). Hussain (2012) found no evidence in English schools to suggest such gaming and found improvement for all students in the schools studied. Furthermore, the improvement in student attainment was found to be maintained in student data for

the following 3 years. These studies all reflect regular school inspections of all schools, looking at the impact of visits and resulting assessments on student attainment. This study particularly analyzes newer, more targeted risk-based models of inspections and how these models affect changes on a broader set of indicators over a period of 3 years.

In this study, we are particularly interested in how school inspections impact on outcomes of schools and how outcomes of schools change over time after schools have been inspected and been placed in different inspection treatment categories. Our study will broaden the previously used definitions of impact by also investigating changes in other (peripheral) subject areas (e.g. geography, biology) and other performance indicators included (to some extent) in the early warning analysis of the inspectorate of education.

We expect schools in the weak and very weak inspection treatment to show more improvements on these indicators compared to schools in the basic category, particularly on indicators that are part of the early warning analysis such as school and exam grades in core subjects (mathematics and literacy). There are a number of arguments that underpin this assumption:

Weak and very weak schools are under increased inspection scrutiny to improve and receive more frequent inspection visits in which their improvement on inspection standards is monitored and targets are set around improvement of key inspection indicators; they face sanctions when failing to improve within 2 to 3 years. Weak and very weak schools also receive inspection feedback and support from other organizations (e.g. external consultants and their school board) while implementing their improvement plan. Moreover, schools in the basic category have less need and room for improvement as they are placed in this inspection category because of already relatively high levels of student attainment. Weak and very weak schools, on the other hand, will have plenty of opportunity to improve as their students are performing below what is expected of them given their entry level.

Several studies suggest that high levels of inspection scrutiny through visits, feedback and consequences for failing have an effect on the improvement of schools, both in raising the school's awareness of relevant standards of educational quality, while also supporting the implementation of improvements. 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, such as those in inspection frameworks and early warning analyses, present the details of what is expected of schools; they create boundaries or domains for attention with respect to educational quality. These standards and, particularly, the threshold to identify failing schools are expected to be important aspects of the impact of school inspections as school officials would select the action that they perceive to have the highest yield, given their planning horizon, budget and appetite for risk. Hanushek and Raymond (2002), for example, found failing schools in the USA on the brink of being sanctioned to make dramatic improvements in the year after identification of these failures. Schools having scores close to the threshold of being sanctioned altered their behaviour more than schools further away from that threshold.

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 are seen as insufficient to motivate schools to perform to high standards (Malen 1999; Elmore and Fuhrman 2001; Nichols et al. 2006). Weak and very weak schools that fail to improve within 2 years after the first assessment can be put forward for financial and/or administrative sanctions by the Ministry of Education, while they are also placed on a list of failing schools on the internet to inform the local community and parents. Penzer (CfBT 2011) explains how such inspection consequences discipline schools to accept unfavourable inspection conclusions.

During inspection visits of weak and very weak schools, inspectors also give feedback about specific weaknesses in the performance of schools on inspection standards, while sometimes also providing advice on how to improve. Such feedback and advice further motivate improvement and inform professional development, particularly when the feedback is clear and explicit, when it is matched by an assessment of weak points as unsatisfactory, when it is provided in a constructive manner and when there is an agreement between an inspector and the school to address the feedback in an improvement plan (such as in weak and very weak schools) (Matthews and Sammons 2004; Dobbelaer et al. 2013; Ehren and Visscher 2008; Gray and Gardner 1999; Erdem and Yaprak 2013; Kelchtermans 2007).

The intensive inspection feedback to weak and very weak schools, their intense monitoring and support for improvement are thus expected to result in more (rapid) improvement of these schools compared to schools in the basic inspection category. The next section will explain the methodology used to study these assumptions.

5 Methodology

The study compared changes in student attainment and additional performance indicators in secondary schools that have been assigned to different inspection treatment categories (basic, weak, very weak) following an early warning analysis in 2011. Multilevel random effect models were used to compare performance of schools in the basic category to schools in the weak/very weak category over a period of 4 years on a set of indicators described below.

5.1 Selection of schools

The schools in our study include a sample of HAVO/VWO schools. A two-stage sampling design was used to select these schools. 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). The results from the early warning analysis in May 2011 were used to select secondary schools from different inspection categories. 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), of the total of 548 Dutch HAVO/VWO schools. 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. In year 1, 17 (6 %) schools were classified as either weak or very weak. In year 2, 12 (4 %) schools were classified as either weak or very weak. In year 3, seven (3 %) schools were classified as either weak or very weak. There are 24 schools that are categorized as either inspection category weak or very weak. Some of these schools are categorized as weak or very weak on more than one occasion; i.e. they are in the same category for either 2 years of the survey or for 3 years.

Table 1 compares schools in the basic and (very) weak inspection categories on a number of performance indicators. A *t* test (comparing schools in the basic category versus schools in the very weak and weak categories) indicated no significant differences on any of the indicators.

The national non-profit organization 'Schoolinfo' provided us with secondary data on the majority of these schools (266, 88 %) at four time points between 2009 and 2013 (the year prior to the early warning analysis, year 1, year 2 and year 3). These data allowed us to measure change in the number of students in the school, parent and student satisfaction, scheduled and taught hours, the number of external evaluations,

Table 1 Performance of schools in basic/weak category

Year 2009–2010 (prior to early warning analysis)	Schools in the basic category	Schools in the weak category	Schools in the very weak category
School size	Mean	Mean	Mean
	810	716	522
Average grade school exam Dutch literacy HAVO[a]	6.41	6.38	6.10
Average grade central exam Dutch literacy HAVO	6.03	5.98	6.15
Average grade school exam Mathematics HAVO	6.23	6.45	6.35
Average grade central exam Mathematics HAVO	6.28	6.50	6.55
Average grade school exam Dutch literacy VWO	6.74	6.73	7.05
Average grade central exam Dutch literacy VWO	6.20	6.28	6.50
Average grade school exam Mathematics VWO	6.42	6.52	6.60
Average grade central exam Mathematics VWO	6.21	6.32	6.20
Throughput lower grades HAVO[b]	96 %	95 %	97 %
Throughput lower grades VWO	96 %	95 %	97 %
Throughput upper grades HAVO	65,7 %	68,0 %	64,9 %
Throughput upper grades VWO	65,7 %	68,0 %	64,9 %
Percentage of students in poverty area	7.01 %	7.32 %	6.92 %
Average student satisfaction (scale of 1–10)	6.68	7.07	5.89
Average parent satisfaction (scale of 1–10)	7.01	7.32	6.92
Scheduled number of teaching hours HAVO	1054	1040	1081
Actual number of taught hours HAVO	1056	975	1015
Scheduled number of teaching hours VWO	1016	1040	1081
Actual number of taught hours VWO	997	961	1026
Percentage sick leave of school staff	4,4 %	4,1 %	.

[a] Grades are given on a 1–10 scale where a '6' is a pass

[b] Percentages throughput represent students not repeating a grade

student achievement on school and central exams, and throughput indicators of lower and upper grades (i.e. number of students progressing without repeating a grade). Data from 2009–2010 acted as a baseline before the early warning analysis of the inspectorate and the assignment of schools to different inspection categories. Table 2 summarizes the available data for the 4 years of data collection.

Unfortunately, Schoolinfo does not have data on all the performance indicators of all the schools for the 4 years of our data collection. Table 3 shows the patterns of missingness within the secondary school data. In the column labelled pattern, '1' represents a response and the '.' represents a missing value. Therefore, the response pattern '.1.' refers to schools that only have responses at one time point, which is the second time point.

5.2 Data analysis

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 analysis indicates if 'basic' versus 'weak/very weak' schools respond differently when faced with different inspection assessments and treatments. Due to the very small number of schools categorized as very weak, this category has been combined with the 'weak' category to create a binary variable of inspection category (basic versus weak/very weak). 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 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 et al. 2009; Thurstone 1934). The analysis is conducted on the predicted scale scores. 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. Here, we only provide detailed reports of significant differences.

6 Results: changes over time in secondary school student attainment

This section describes changes in student achievement as measured in school exams and in national standardized exams in a number of different subjects (Dutch literacy, mathematics A and B, geography, chemistry, biology, economics A and B) by academic track (VWO or HAVO). We particularly looked at whether schools in different inspection treatment categories (weak/very weak versus basic) have different processes of change after having been placed in an inspection category and whether there is a difference in the level of change of core subjects versus other subjects. A graphical display of the results of the hierarchical linear modelling (HLM) is provided followed by a table containing the coefficients from the HLM models. The results for the subject grades are divided into school and central exam grades and by whether the exam grades

Table 2 Overview of data collection

Year 2009–2010	Year 2010–2011	Year 2011–2012	Year 2012–2013
• Number of students 2009–2010	• Number of students 2010–2011	• Number of students 2011 2012	• Number of students 2012 2013
• Number of students in exams per profile 2009–2010	• Number of students in exams per profile 2010–2011	• Number of students in exams per profile 2011–2012	• Number of students 2012–2013 in APC (poverty area)
• Average grade school exam per subject 2009–2010[31]	• Average grade school exam per subject 2010–2011	• Average grade school exam per subject 2011–2012	• Number of students 2012–2013 NOT in APC (poverty area)
• Average grade central exam per subject 2009–2010	• Average grade central exam per subject 2010–2011	• Average grade central exam per subject 2011–2012	• Percentage of students 2012–2013 in APC (poverty area)
• Throughput lower grades 2009–2010[42]	• Throughput lower grades 2010–2011	• Throughput lower grades 2011–2012	• Percentage of students 2012–2013 NOT in APC (poverty area)
• Throughput upper grades 2009–2010	• Throughput upper grades 2010–2011	• Throughput upper grades 2011–2012	• Average student satisfaction 2012–2013
• Number of students 2009–2010 in APC (poverty area)	• Number of students 2010–2011 in APC (poverty area)	• Number of students 2011–2012 in APC (poverty area)	• Average parent satisfaction 2012 2013
• Number of students 2009–2010 NOT in APC (poverty area)	• Number of students 2010–2011 NOT in APC (poverty area)	• Number of students 2011–2012 NOT in APC (poverty area)	• Number of external evaluations 2012–2013
• Percentage of students 2009–2010 in APC (poverty area)	• Percentage of students 2010–2011 in APC (poverty area)	• Percentage of students 2011–2012 in APC (poverty area)	
• Percentage of students 2009–2010 NOT in APC (poverty area)	• Percentage of students 2010–2011 NOT in APC (poverty area)	• Percentage of students 2011–2012 NOT in APC (poverty area)	
• Average student satisfaction 2009–2010	• Average student satisfaction 2010–2011	• Average student satisfaction 2011–2012	
• Average parent satisfaction 2009–2010	• Average parent satisfaction 2010–2011	• Average parent satisfaction 2011–2012	
• Scheduled teaching hours per department (HAVO/VWO) per year 2009–2010	• Scheduled teaching hours per department (HAVO/VWO) per year 2010–2011	• Scheduled teaching hours per department (HAVO/VWO) per year 2011–2012	
• Actual taught hours per department (HAVO/VWO) per year 2009–2010	• Actual taught hours per department (HAVO/VWO) per year 2010–2011	• Actual taught hours per department (HAVO/VWO) per year 2011–2012	
• Percentage of staff on sick leave 2009	• Percentage of staff on sick leave 2010	• Number of external evaluations 2011–2012	
		• Percentage of staff on sick leave 2011	

[31]Subjects included in the study are Dutch literacy, Mathematics (A and B), Geography, Chemistry, Biology, Economics A and B).
[42]Throughput represents percentage of students not repeating a grade.

relate to the VWO track (pre-university education track) or the HAVO track (general higher education track).

Table 3 The pattern of missingness for additional data of secondary schools

Freq.	Percent	Cum.	Pattern
266	88.37	88.37	1111
13	4.32	92.69	.1.
3	1	93.69	.11
3	1	94.68	.11.
3	1	95.68	1...
2	0.66	96.35	...1
2	0.66	97.01	.1.
2	0.66	97.67	.111
2	0.66	98.34	11.
5	1.66	100	(other patterns)
301	100	XXXX	

The results indicate that changes in Dutch literacy grades differed by inspection category, particularly for the average school exam grade in Dutch for VWO and for the central exam grade in Dutch for HAVO (Table 4).

As shown in Fig. 1, students in the VWO academic track in schools in the weak/very weak category (estimate = 6.85, SE = 0.05) initially scored significantly higher on the *school exams* compared to students in schools in the basic category of inspection (estimate = 6.74, SE = 0.02). However, the scores declined on the school exams in Dutch literacy over the 3 years so that in year 3, they scored on average less than students in schools in the basic category of inspection. Students in schools in the basic category of inspection did not show the same decline in scores over time. The differences between the inspection categories were significantly different in year 3 for students in the VWO track, with students in schools in the weak/very weak inspection category (estimate = 6.56, SE = 0.07) scoring significantly lower than students in schools in the basic inspection category (estimate = 6.56, SE = 0.02). Indeed, there was a significant interaction between school exam grades in Dutch literacy for students in the VWO track (chi2(2) = 7.72, $p < 0.05$). A similar pattern, although much less pronounced, was displayed for the students in the HAVO track, but these differences were not statistically significant (chi2(2) = 1.79, $p > 0.05$).

There is also some evidence, shown in Fig. 2, that scores on the average *central exam grade* in Dutch differed by inspection category over time. There is a significant interaction between changes in the central exam grades over time and inspection category of the school for students in the HAVO academic track (chi2(2) = 6.75, $p < 0.05$). Students in schools in inspection category weak/very weak in the HAVO track (estimate = 5.90, SE = 0.05) initially have lower scores on the central exam than students in schools in the basic inspection category (estimate = 6.02, SE = 0.01). However, between year 1 and year 2, this changes and students in schools in inspection category weak/very weak have higher scores (estimate = 6.24, SE = 0.06) than students in schools in the basic inspection category (estimate = 6.15, 0.02). By year 3, there is little discernible difference

Table 4 Testing the association between school inspection category and changes in Dutch language and literature grades over time

	Est	SE	*Z*	*p* value	95 % Confidence intervals	
Dutch school VWO						
Year						
Year 1	Reference					
Year 2	−0.05	0.01	−3.52	0.00	−0.07	−0.02
Year 3	−0.08	0.01	−5.98	0.00	−0.11	−0.05
Inspection category						
Basic	Reference					
Weak/very weak	0.10	0.05	2.16	0.03	0.01	0.20
Year × inspection category						
Year 2 × weak	−0.12	0.06	−1.95	0.05	−0.25	0.00
Year 3 × weak	−0.21	0.08	−2.54	0.01	−0.37	−0.05
Intercept	6.74	0.02	437.47	0.00	6.71	6.77
sigma_u	0.21					
sigma_e	0.15					
rho	0.65					
Test of interaction	chi2(2) = 7.72, $p = 0.021$					
Dutch school HAVO						
Year						
Year 1	Reference					
Year 2	−0.03	0.01	−2.82	0.01	−0.06	−0.01
Year 3	−0.02	0.01	−1.22	0.22	−0.04	0.01
Inspection category						
Basic	Reference					
Weak/very weak	0.03	0.04	0.64	0.52	−0.06	0.11
Year × inspection category						
Year 2 × weak/very weak	−0.07	0.06	−1.28	0.20	−0.19	0.04
Year 3 × weak/very weak	−0.06	0.07	−0.81	0.42	−0.20	0.08
Intercept	6.40	0.01	488.37	0.00	6.37	6.42
sigma_u	0.17					
sigma_e	0.14					
rho	0.59					
Test of interaction	chi2(2) = 1.79, $p = 0.409$					
Dutch central VWO						
Year						
Year 1	Reference					
Year 2	0.05	0.02	2.45	0.01	0.01	0.08
Year 3	0.23	0.02	11.89	0.00	0.19	0.26
Inspection category						
Basic	Reference					
Weak/very weak	−0.07	0.06	−1.07	0.28	−0.19	0.06

Table 4 (continued)

	Est	SE	*Z*	*p* value	95 % Confidence intervals	
Year × inspection category						
Year 2 × weak/very weak	0.18	0.09	2.02	0.04	0.01	0.35
Year 3 × weak/very weak	0.04	0.11	0.40	0.69	−0.17	0.26
Intercept	6.20	0.02	363.37	0.00	6.17	6.23
sigma_u	0.18					
sigma_e	0.22					
rho	0.40					
Test of interaction	chi2(2) = 4.19, $p = 0.122$					
Dutch central HAVO						
Year						
Year 1	Reference					
Year 2	0.14	0.02	8.35	0.00	0.11	0.17
Year 3	0.20	0.02	12.44	0.00	0.17	0.24
Inspection category						
Basic	Reference					
Weak/very weak	−0.11	0.05	−2.02	0.04	−0.22	0.00
Year × inspection category						
Year 2 × weak/very weak	0.20	0.08	2.55	0.01	0.05	0.35
Year 3 × weak/very weak	0.13	0.10	1.37	0.17	−0.06	0.32
Intercept	6.02	0.01	414.03	0.00	5.99	6.05
sigma_u	0.15					
sigma_e	0.19					
rho	0.38					
Test of interaction	chi2(2) = 6.75, $p = 0.034$					

There is no evidence of differential changes in the maths A and B, geography, chemistry, biology and economics grades on school or on central exams over time by inspection category.

between schools in the different inspection categories (weak/very weak estimate = 6.24, SE = 0.08; basic estimate = 6.22, SE = 0.01). There is also no significant interaction between inspection category and year of the study for the average central exam grades in the HAVO track. A similar patterning of results was found for students in the VWO track, but the differences between the inspection categories were smaller at year 1 (weak/very weak estimate = 6.23, SE = 0.16; basic estimate = 6.20, SE = 0.02), and the interaction over time was not statistically significant (chi2(2) = 4.19, $p = 0.122$).

7 Results: changes over time in other performance indicators

Finally, we looked at changes in other indicators, such as student/parent satisfaction, numbers and ratios of students to staff as outlined in Table 2. The results

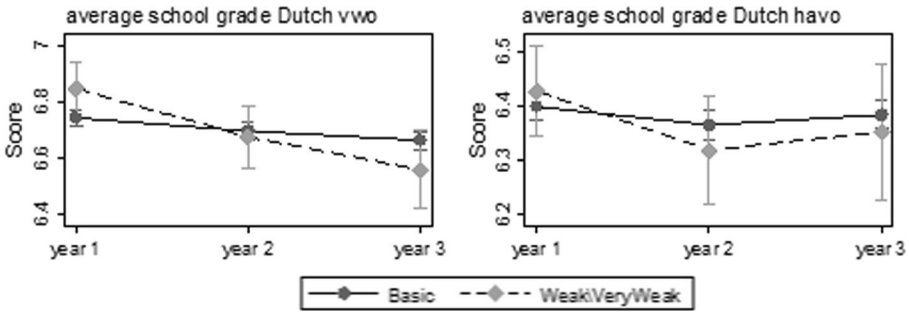

Fig. 1 Changes in average school exam grade in Dutch by inspection category

indicate that schools in the two inspection categories particularly differ in their changes in student satisfaction scores, the number of students per full-time employee and the number of students in the school over time. There is no evidence of differential changes in parental satisfaction, the ratio of students to management full-time employees, the ratio of students to teacher full-time employees, the proportion of students living in poverty areas and the proportion of sick leave days over time by inspection category.

7.1 Student satisfaction

Student satisfaction scores changed differently by inspection category over time (chi2(2) = 6.20, $p < 0.05$). Initially, student satisfaction scores are very similar in the inspection categories (weak/very weak estimate = 6.79, SE = 0.10; basic estimate = 6.78, SE = 0.03). By year 2, students in schools that are in inspection category weak/very weak report significantly lower satisfaction than students in schools in the basic category of inspection (weak/very weak estimate = 6.46, SE = 0.12; basic estimate = 6.75, SE = 0.03). Between years 2 and 3, the scores for student satisfaction for students in school classified as weak/very weak increased significantly (weak/very weak estimate = 7.04, SE = 0.23), whereas satisfaction scores for students in schools in the basic inspection category only slightly increased (basic estimate = 6.83, SE = 0.03) (Table 5; Fig. 3).

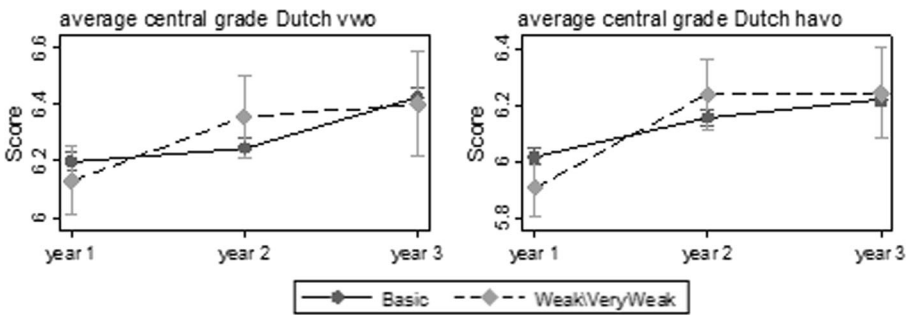

Fig. 2 Changes in average central exam grade in Dutch by inspection category

Table 5 Testing the association between school inspection category and changes in student satisfaction over time

Student satisfaction	Estimate	SE	*Z*	*p* value	95 % Confidence intervals	
Time						
Year 1	Reference					
Year 2	−0.03	0.03	−1.01	0.31	−0.09	0.03
Year 3	0.04	0.03	1.39	0.17	−0.02	0.10
Inspection category						
Basic	Reference					
Weak/very weak	0.01	0.11	0.10	0.92	−0.20	0.22
Time × inspection						
Year 2 × weak	−0.30	0.16	−1.91	0.06	−0.60	0.01
Year 3 × weak	0.20	0.24	0.84	0.40	−0.27	0.68
Intercept	6.78	0.03	251.49	0.00	6.72	6.83
sigma_u	0.28					
sigma_e	0.27					
rho	0.51					

Test of interaction between time and inspection category chi2(2) = 6.20, $p = 0.045$

7.2 Number of students per full-time employee

In year 1 and year 2, the number of students per full-time employee is very similar in schools in both inspection categories. There is, however, evidence that in year 3, the number of students per full-time employee was significantly lower in schools in the weak/very weak inspection category compared to schools in the basic inspection category. The actual average difference between the inspection category groups is approximately 0.5 (weak/very weak estimate = 11.03, SE = 0.23; basic estimate = 11.55, SE = 0.08), which means that, in schools that are in the inspection category weak/very

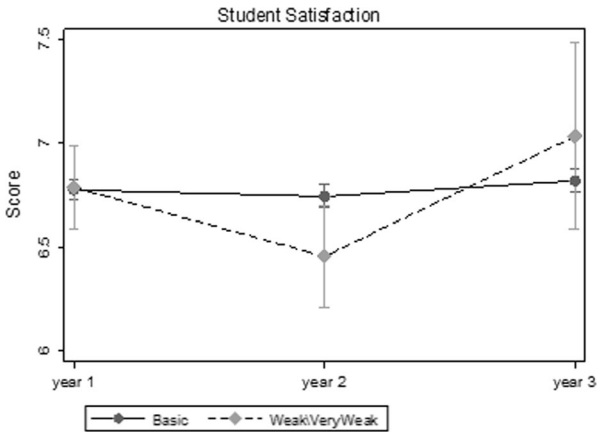

Fig. 3 Changes in student satisfaction by inspection category

Table 6 Testing the association between school inspection category and changes in the ratio of students per full-time employee over time

Students per FTE	Estimate	SE	*Z*	*p* value	95 % Confidence intervals	
Time						
Year 1	Reference					
Year 2	0.36	0.04	8.43	0.00	0.28	0.44
Year 3	0.63	0.04	14.90	0.00	0.55	0.71
Inspection category						
Basic	Reference					
Weak/very weak	0.02	0.16	0.15	0.88	−0.29	0.34
Time × inspection						
Year 2 × weak	0.01	0.21	0.06	0.95	−0.39	0.42
Year 3 × weak	−0.55	0.26	−2.12	0.03	−1.05	−0.04
Intercept	10.92	0.08	134.31	0.00	10.76	11.08
sigma_u	1.27					
sigma_e	0.47					
rho	0.88					

Test of interaction between time and inspection category chi2(2) = 5.30, $p = 0.071$

weak, there was approximately half a pupil less per full-time employee on average compared to schools in inspection category basic (Table 6; Fig. 4).

7.3 Total number of students

There is also a significant interaction between the number of students in the school over time and the inspection category of the school (chi2(2) = 15.86, $p < 0.05$). There is very little change in the number of students in the school over time for schools in the basic inspection category (year 1 estimate = 784.5, SE = 19.3; year 2 estimate = 788.1, SE = 19.3; year 3 estimate = 787.7, SE = 19.3). However, schools in the weak/very

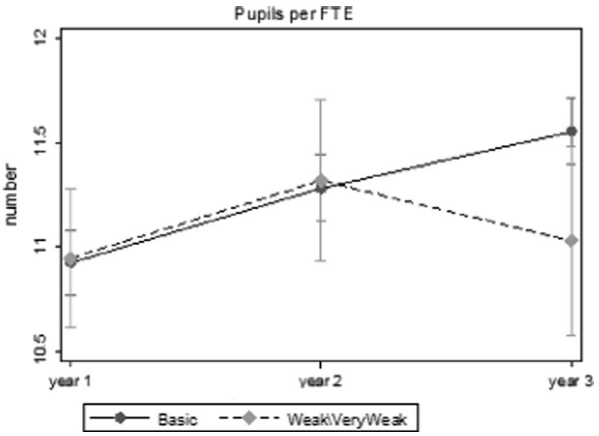

Fig. 4 Changes in students per FTE by inspection category

Table 7 Testing the association between school inspection category and changes in the number of students in the school over time

Number of students in school	Estimate	SE	*Z*	*p* value	95 % Confidence intervals	
Time						
Year 1	Reference					
Year 2	3.59	4.58	0.78	0.43	−5.38	12.56
Year 3	3.17	4.58	0.69	0.49	−5.80	12.14
Inspection category						
Basis	Reference					
Zwak/zeer zwak	39.37	17.53	2.25	0.03	5.01	73.73
Time × inspection						
Year 2 × zwak	−16.70	21.68	−0.77	0.44	−59.18	25.79
Year 3 × zwak	−110.67	28.19	−3.93	0.00	−165.92	−55.41
Intercept	784.52	19.28	40.69	0.00	746.73	822.31
sigma_u	320.32					
sigma_e	51.53					
rho	0.97					

Test of interaction between time and inspection category chi2(2) = 15.86, $p = 0.000$

weak inspection category start with a significantly higher number of students than schools in the basic category (39 more students on average). In year 2, there are no differences in the number of students by inspection category, but by year 3, schools in inspection category weak/very weak have fewer students (71 fewer on average) than schools in the basic inspection category. Changes in student to staff ratios are either the result of taking on more staff or taking on fewer students. The results here suggest that there are fewer students attending schools that are in the weak/very weak inspection category. As parents are free to choose a school, they may have opted out of the weak/very weak schools and entered their child in a higher performing school. Alternatively, schools may have enrolled fewer students to make better use of their resources and staff and improve performance. Unfortunately, the mechanism cannot be tested with these data (Table 7; Fig. 5).

8 Conclusion and discussion

Many inspectorates of education have recently moved towards more proportional risk-based inspection models, targeting high-risk schools for visits, while schools with satisfactory student attainment levels are excluded from inspections. The Bratislava memorandum of the European Association of Inspectorates of Education (SICI) even mentions such models as a way forward towards more agile methods of school inspections. This paper looks into these newer inspection models and aims to enhance our understanding of the potential effectiveness of such targeted models on student attainment and other performance indicators. We particularly looked into the impact of the introduction of risk-based inspections in the Netherlands in 2008, analyzing

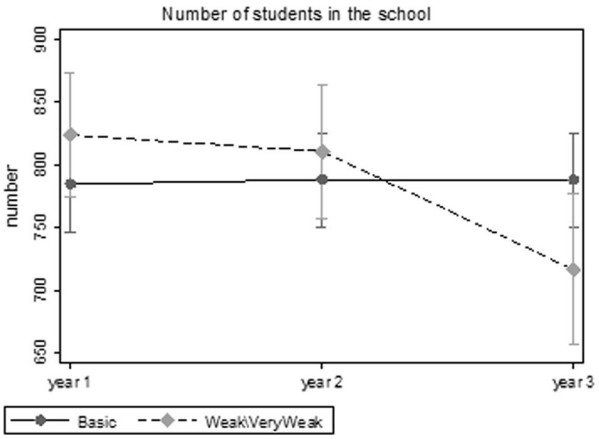

Fig. 5 Changes in the number of students in the school by inspection category

changes in schools that are placed in different inspection treatment categories after the early warning analysis. Using data from the non-for-profit organization Schoolinfo on students' attainment in a number of subjects as well as in other indicators (e.g. student/parent satisfaction), we analyzed whether there are differences in how schools in different inspection categories (basic versus weak/very weak) change over a period of 3 years and whether there are differences in changes on the indicators that are part of the early warning analysis (student attainment and student numbers) and indicators excluded from the framework (e.g. student/parent satisfaction).

As school inspections have been in place for many years, albeit in different guises and with different foci, our findings are exploratory at best and do not allow for strong causal conclusions about the overall effect of inspections. Even though we analyzed changes in schools before and after the introduction of a new risk-based inspection model, many schools will have known about upcoming changes and will have prepared themselves for these changes. Such preparations may have diluted our findings in preventing us from finding significant differences in changes between weak/very weak schools and those in the basic inspection category. The small number of weak and very weak schools in our study will also limit the number of significant differences that we will find in performance of these schools compared to schools in the basic category, potentially underestimating the actual effect of risk-based inspections. Our models, however, do allow for an exploration of the differences in improvement trajectories between the two groups of schools on a range of indicators (both in and outside of the inspection framework) over a short period of time. Such findings provide insights into the types of changes in schools following targeted inspections.

8.1 Changes in student attainment of schools in different inspection categories

The results indicate differences in changes in student achievement results in secondary schools, but only in Dutch literacy. These differences become more prominent over time and are particularly significant in the second and third years after the early warning analysis, suggesting that it takes 2 to 3 years for inspection visits to have an impact on student attainment. Weak/very weak schools in the VWO track showed a decline in

school exam results in Dutch language compared to (stable) scores in the VWO schools in the basic category, whereas weak/very weak schools in the HAVO track show increasing scores on the central exam results in Dutch language compared to schools in the basic category.

The decline in school exam scores may be explained by schools marking their own school exams stricter, following more strict guidelines of the inspectorate. A previous study by De Lange and Dronkers (2007) showed a large gap between students' scores on school and central exams, suggesting that schools try to improve their pass rates and their position in league tables by boosting students' grades on the school exams. As a result, the inspectorate of education added an indicator to their assessment framework, which restricts the difference between the two scores. This may have resulted in weak/very weak schools marking their own school exams stricter, which would lead to a decline in scores.

Another explanation is the unstandardized nature of school exams and potential unreliable marking, which may have resulted in fluctuation of and a potential decline of results. However, the availability of marking scripts, as well as second marking of school exams, suggests that there is a high level of scrutiny of these exams, which would prevent error in measuring differences between schools in the two inspection categories, as well as incorrect measurement of changes in school performance over time.

8.2 Changes in other performance indicators in schools in different inspection categories

We also analyzed differential changes in other performance indicators between schools in different inspection categories, expecting the most changes in weak/very weak schools on indicators part of the early warning analysis of the inspectorate of education (student numbers). Our results only show differences in changes in student satisfaction, student-staff ratios and number of full-time students in weak/very weak schools compared to schools in the basic inspection category, suggesting that schools do not specifically target improvement on indicators in the early warning analysis. Student satisfaction declined, as well as student numbers and student-staff ratios in weak and very weak schools over time. This would suggest that students are less likely to choose schools that are evaluated as weak or very weak by the inspectorate, and students in weak and very weak schools become less satisfied when the school is assessed to be failing. Such a decline in student satisfaction may result from an overall lack of morale in the school, following a negative inspection assessment, or may be caused by the fact that time and effort go into raising student achievement in core subjects to the disadvantage of other subjects or activities in the school, creating a range of side effects such as lack of morale for long-term improvement or narrowing the curriculum to tested subjects and content.

The improved scores on the central examination in literacy only 2 years after the inspection visit raise further concerns about potential unintended consequences of risk-based inspections. van den Berg and Vandenberghe (1981), Stringfield (2002) and Visscher (2002) found that complex change generally takes 5 to 10 years to result in higher student outcomes, suggesting that the improvement activities in our study schools were potentially aimed at short-term gains, implemented under high pressure

to show speedy results. The fact that the timeline of improved scores coincides with the inspection consequences for failing schools, where sanctions kick in 2 to 3 years after their categorization as weak or very weak, supports such a claim.

These results lead us to question whether risk-based inspection models are the best way forward. This takes us back to our initial reflection on drawbacks of risk-based inspections and whether the early warning methodology allows for a detection of high-risk schools.

De Wolf and Honingh (2014) suggest that risk-based models are best placed in settings where risks are actually known and measurable and where inspectorates are capable of taking into account the level of uncertainty that comes with using a risk-based model. It is questionable whether risks of declining quality are actually measurable in education, given the multiple causes of decline and the great variety of contexts in which schools operate and that affect their performance. Risks are particularly unpredictable in settings of large variety in the inspected organization, such as schools with different student populations in different urban and rural contexts, where services are delivered in interaction of different providers (e.g. teachers, school boards, after school care organizations responsible for aspects of teaching and the school's organization) and when there is limited data available to inform an early warning analysis. In such situations, accurate prediction of risks is difficult according to De Wolf and Honingh (2014) and is often not more than a best guess. As Ehren and Honingh (2011) outline, the focus on student achievement in the Dutch early warning analysis does not actually allow for an analysis of risks as the performance of schools is already failing when they are identified for targeted inspections; inspection visits become a remedy for failing schools instead of a model to prevent such failure and improve the performance of all schools in the education system.

Power (1999) also suggests that auditing labels, such as the ones from an early warning analysis of schools with/without risks, do not invite or provoke public dialogue as they do not contribute to the empowerment of external parties, such as parents or students. The labels of schools in different inspection categories and lack of reports of schools in the 'basic category' limit the opportunities of parents, students and other stakeholders external to the school to enter into communication and dialogue with the school about further improvements. The model of early warning analysis excludes a large group of schools in the basic inspection category from feedback and conversations about further improvement and implies a shift from a highly interactive inspection to a primary paper-based audit exercise, assuming that such interactive and broad-based inspections and evaluations are now organized by schools in their horizontal accountability arrangements. Several authors have argued that such models of external audits and checks of internal systems for self-inspection are the best chance of effective regulation (Braithwaite 1989), but others particularly highlight the 'ritualization' of evaluations where the focus is shifted from the evaluation of the 'quality of one's deeds' to the 'quality of one's records' (Power 1999).

Power (1999) has also pointed to potential goal displacement of audits when auditing becomes an organizational ritual and loses its relevance of having an impact on the objects that it evaluates and controls. Such goal displacement is particularly relevant in the context of risk-based inspections where the definition of high-quality teaching and learning is narrowed down to a small number of indicators used in the early warning analysis. Examples of such goal displacement were highlighted by De

Wolf and Honingh (2014) who found that the use of risk-based models lowered expectations of Dutch schools that now feel that having no risks equals good performance. This conclusion has already led the Dutch Inspectorate of Education to revisit their inspection model and to increase the intensity of their inspections of schools in the basic category, introducing more differentiated targets for these schools and providing all these schools with regular feedback on their performance and sharing good practises of high-performing schools across the system. As such models are more costly and require more visits to schools, it remains to be seen whether other countries that have introduced risk-based, proportionate models will follow suit.

Acknowledgments This research was funded by the EU Lifelong Learning Programme, grant number: 511490-LLP-1-2010-1-NL-KA1-KA1SCR.

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.

Author (2012). Risk-based school inspections of Dutch schools and school boards: a critical reflection on intended effects and causal mechanisms. http://schoolinspections.eu/impact/wp-content/uploads/downloads/2012/05/Netherlands_PT.pdf. Retrieved Jan 2016.

Béguin, A., & Ehren, M. (2011). Aspects of accountability and assessment in the Netherlands. *Zeitschrift für Erziehungswissenschaft, 14*, 25–36.

Braithwaite, J. (1989). *Crime, shame and reintegration*. Cambridge: University press.

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

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.

De Wolf, I. F., & Honingh, M. (2014). Risicogestuurd toezicht niet vrij van risico's [risk-based inspections are not free of risks]. In F. Mertens, J. Scherpenisse, & M. van der Steen (Eds.), *Reflecties op de ontwikkeling en professionalisering van het toezicht; 10 jaar Leeratelier Toezicht en Naleving [Reflections on the development and professionalization of inspections; 10 years of teaching on the inspection and compliance programme]* (pp. 45–59). Den Haag: NSOB.

DiStefano, C., Zhu, M., & 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.

De Lange, M., & Dronkers, J. (2007). *Hoe gelijkwaardig blijft het examen tussen scholen*? *Discrepantie tussen de cijfers voor het schoolonderzoek en het centraal examen in het voortgezet onderwijs tussen 1998 en 2005.* [The equivalence between schools of the Dutch secondary final examination. Discrepancies between the grading of the central and school part of the final examinations of secondary education between 1998 and 2005]. European University Institute working paper EUI SPS 2007/3. Florence: EUI.

Ehren, M. C. M., & Honingh, M. (2011). Risk-based school inspections in the Netherlands: a critical reflection on intended effects and causal mechanisms. *Studies in Educational Evaluation (special issue), 37*(4), 239–248.

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.

Elmore, R. F., & Fuhrman, S. H. (2001). Research finds the false assumption of accountability. *Phi Delta Kappan, 67*(4), 9–14.

ERDEM, A. R., & YAPRAK, M. (2013). The problems that the classroom teachers working in villages and county towns confront in educational inspection and their opinions concerning the effect of these problems on their performance. *Educational Research and Reviews, 8*, 455–461.

EP-Nuffic (2015). *Education System the Netherlands*; *the Dutch Education system described*. The Hague: Nuffic. https://www.epnuffic.nl/en/publications/education-system-the-netherlands.pdf. Retrieved Jan 2016.

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 Jul 2014

Gray, C., & Gardner, J. (1999). The impact of school inspections. *Oxford Review of Education, 25*(4), 455–469.

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.

Hussain, I. (2012). *Subjective performance in the public sector: evidence from school inspections*. London School of Economics and Political Science. Centre for Economic Performance.

Inspectie van het Onderwijs (2012). Inspectorate profile: the Netherlands. http://www.sici-inspectorates.eu/getattachment/f18c145e-4373-4405-8ab7-39b715b5d1e5

Kelchtermans, G. (2007). Macropolitics caught up in micropolitics: the case of the policy on quality control in Flanders (Belgium). *Journal of Education Policy, 22*(4), 471–491.

Klerks, M. (2013). The effect of school inspections: a systematic review. www.schoolinspections.eu. Retrieved Jan 2014.

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: Institute of Education.

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.

Penzer, G. & CFBT EDUCATION TRUST. 2011. *School inspections what happens next*? [Online]. Reading: CfBT Education Trust.

Power, M. (1999). *The audit society: rituals of verification*. Osford: OUP.

Quene, H., & 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.

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.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration, 18*(2–3), 277–310.

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

Stringfield, S. (2002). Big change questions: is large-scale educational reform possible? *Journal of Educational Change, 3*, 63–73.

Thurstone, L. L. (1934). 'The vectors of mind'. *Psychological Review, 41*, 1–32.

van den Berg, R., & Vandenberghe, R. (1981). *Onderwijsinnovatie in een verschuivend perspectief* [Shifting perspectives on educational innovation]. Tilburg: Uitgeverij Zwijsen bv.

Visscher, A. J. (2002). A framework for studying school performance feedback systems. In A. J. Visscher & R. Coe (Eds.), *School improvement through performance feedback* (pp. 41–75). Lisse: Swets & Zeitlinger.