Educ Asse Eval Acc (2015) 27:323–346
DOI 10.1007/s11092-015-9212-y

Risk-based educational accountability in Dutch primary education

A. C. Timmermans · I. F. de Wolf · R. J. Bosker · S. Doolaard

Received: 11 March 2014 / Accepted: 7 January 2015 / Published online: 30 January 2015

Abstract A recent development in educational accountability is a risk-based approach, in which intensity and frequency of school inspections vary across schools to make educational accountability more efficient and effective by enabling inspectorates to focus on organizations at risk. Characteristics relevant in predicting which schools are "at risk on adverse effects" and robustness of results of risk-based analyses over multiple cohorts were assessed by an empirical analysis of Dutch primary schools. Adverse effects were defined as below average final achievement and/or below average value added. School composition, previous underperformance, insufficient judgments on having a systematic evaluation approach, evaluation of support, and monitoring student performance appeared as factors related to subsequent underperformance of schools. Although a rich set of possible risk factors was available, further investigation of a large number of schools is required in order to find nearly all underperforming schools. However, a group of about 40 % of the schools showed very small risk on underperformance, which represents the efficiency gain when risk-based school accountability would be applied. Furthermore, whether schools are (in)accurately classified in the risk analysis as "at-risk" schools depends heavily on the chosen caesura.

Keywords Educational accountability · Risk assessment · Primary education · Value added

1 Introduction

Most European countries have Inspectorates of Education to assess the quality of public schools. In general, the aims of these inspectorates are to guarantee a

A. C. Timmermans (✉) · R. J. Bosker · S. Doolaard
GION Education/Research, University of Groningen, Grote Rozenstraat 3, 9712 TG Groningen, The Netherlands
e-mail: A.C.Timmermans@rug.nl

I. F. de Wolf
Dutch Inspectorate of Education, Utrecht, The Netherlands

minimum quality level of education and to improve the schools' quality. The main instruments of the inspectorates are school inspections, during which inspectors assess the quality of the educational processes within the schools and try to identify underperforming schools. Besides school inspections, many countries use school accountability systems in which value-added or related performance indicators are used as quantitative measures to identify underperforming schools (Figlio and Loeb 2011; Inspectie van het Onderwijs 2009a, 2011; Ofsted 2010, 2011). Recently, some of the European inspectorates introduced a risk-based strategy to improve the effectiveness and efficiency of the school inspections (De Wolf and Verkroost 2011; Inspectie Onderwijs van het 2009b; Ofsted 2011). The transition to risk-based school inspection systems became possible when standardized data at the school level became available to identify schools that are at risk of underperformance.

A risk-based inspection system is assumed to be more *effective*, because it enables the inspectorates to focus on the organizations (schools) at risk (Sparrow 2000). These are schools with a higher risk of noncompliance or under- performance, which may benefit more from inspections than well-performing organizations. Risk-based inspections are also assumed to be more *efficient* (less time consuming) than the traditional inspection systems. This efficiency gain lies in a less intensive inspection regime for well-performing organizations. However, whether or not a risk-based inspection strategy may actually lead to an effectiveness and efficiency gain depends on the answer to the following question: "How well can we estimate which schools are at risk?" The better the estimate of which schools are at risk, the better it is possible to target the inspections at those schools who benefit most from the inspections and to introduce a less intensive inspection regime for schools not at risk. In this study, we explored a methodology to detect which schools are "at risk of underperformance" for the purpose of risk-based educational accountability in Dutch primary education. To that end, we will start with a description of risk analysis, followed by a short description of the Dutch primary education system.

1.1 Risk analysis

Risk analysis can be defined as the "systematic use of available information to determine how often specified events may occur and the magnitude of their consequences" and has its roots in probability theory (Molak 1997; Standards Association of Australia 1999). Numerous of these kinds of risk analyses exist in other research traditions, which are mostly applied to determine safety risks and risks in project management (e.g., Keeney and Von Winterfeldt 2011; De Jong 2012). In risk analysis, a risk can be defined as the probability for some adverse effect to occur given some conditions (risk factors). The definition of an adverse effect is therefore a crucial first step in risk analysis (Standards Association of Australia 1999); however, it is often considered to be a value judgment (Molak 1997). Examples of well-known adverse effects within the tradition of risk analysis are death, diseases, failure of nuclear power plants, and loss of investments. For the specific context of educational accountability in primary education, we can think of several adverse effects. However, in this study, we investigated two adverse effects.

The first one is low student performance at the end of a formal stage of education at a school (low final achievement). A second adverse effect relates to a limited progress in students' performance during the formal stages of education at a school (low value added).

After defining adverse effects, the factors related to the adverse effects have to be determined, as well as the levels of these factors for which the adverse effect becomes more prevalent. In the context of risk-based educational accountability, this step in the risk analyses would be to answer the following question: "What characteristics of schools are relevant in predicting at which schools students have low performances or make limited progress during a formal stage of schooling?" Two methods of risk analysis are common: estimating risks via empirical methods and by using the judgments of experts in the field (Molak 1997). In the current study, we focus on an empirical method of risk analyses to assess which characteristics of schools were associated with the low performance and low progress in performance in schools in Dutch primary education.

In the empirical methods, it is assumed that one can establish the probability of occurrence of adverse effects on the basis of historical data (Molak 1997). Regression models on retrospective data are therefore frequently used in risk analysis. In these models, occurrence of an adverse effect at the current time point (t) is estimated based on all information available at previous time points ($t-1$ and before). For example, the previous performance of schools ($t-1$ and before) might be a predictor of the current performance of schools (t). Of course, there may be many more factors at time point $t-1$ that predict the current performance of schools. Examples of such other factors are school size, the schools' student population, and the experience of the teaching staff. In the empirical analysis, it is then assumed that the factors which predict the occurrence of adverse effects at the current time point (t) will also predict this adverse effect at future time points ($t+1$). This could, for example, imply that the current performance of schools (t) is a predictor of the future performance of schools ($t+1$) in a similar way as the previous performance ($t-1$ and before) predicted the current performance (t). Based on this assessment, it can be established to which extent the prevalence of adverse effects can be accurately predicted, or in other words how well it is possible to establish which schools are at risk.

In order to estimate which schools are at risk, standardized data at the school level are crucial in order to predict adverse effects. In most countries, the school-level data available include information on test scores, school performance indicators, student and teacher characteristics, signals concerning children's safety within schools as well as school practices and policies. In several educational accountability systems, raw performance indicators are available as a measure of the final achievement of students in schools and value-added models have increasingly been adopted to assess the progress of students during a particular period of schooling (e.g., Betebenner 2007, 2009; Ofsted 2010; Ray 2006; Sanders 2003; Sanders and Horn 1994). Together, these two performance indicators cover the adverse effects as defined in this particular study. Other available information, such as student and teacher characteristics, signals concerning children's safety within schools, and school practices and policies can be used to predict these adverse effects.

1.2 Risk-based educational accountability in Dutch primary education

Dutch primary education is intended for all children from approximately age 4 (pre-kindergarten) up to and including age 12 (grade 6). In 2013, approximately 1,500,000 pupils were enrolled in 6500 primary schools (Ministry of Education 2014). During primary education, many schools monitor the progress of their student by means of the so-called monitoring and evaluation systems. These systems consist of a series of tests, which are administered by schools, mostly for their own use. In the final grade, about 85 % of the schools administer the "school leavers test." The school leavers test consists of the basic subjects and was designed to help teachers to formulate a track recommendation for secondary education. The results on the school leavers test, however, are also used to estimate school performance indicators in the primary education accountability system.

The accountability framework for Dutch primary education contains five school performance indicators (Inspectie van het Onderwijs 2011). The performance indicators are based on the results of students on tests at the end of primary education (mostly the school leavers test), tests in monitoring systems during primary education and the amount of grade retention. To pursue fair comparisons of the performance of schools at the end of primary education, a comparison is made between a schools' scores on a test and the results of schools with a similar student population. This latter is based on the percentage of students within schools with lowly educated parents. Next to the performance indicators, school processes, policy, and social outcomes are assessed during school inspections.

A schematic overview of the risk-based inspection strategy of the Dutch Inspectorate of Education is given in Fig. 1. The first step in this process is gathering information on outcomes (school performance indicators), annual accounts, and failure signals. The annual accounts pertain to school-level data on staff (turnover), pupils, and the financial situation. Failure signals include, for example, complaints lodged by parents or media reports. The second step in the process is a risk analysis based on this information. This risk analysis is usually conducted once a year, but failure signals of schools may ask for a risk analysis on other occasions. If the analysis does not reveal any risks, the Inspectorate has sufficient trust in the quality of the educational process to qualify the school for the basic inspection program: a school inspection once every 4 years. When the risk analysis reveals possible risks, additional information is requested from the school (quality study). The nature of and the background to the alleged risks are then investigated and a more extensive analysis of the collected data is conducted. If eventually all appears to be in order, the school is placed in the basic inspection program as well. When a school is found to have shortcomings with regard to quality, it will receive tailored inspection for weak or unsatisfactory quality. For a more detailed description of the risk-based strategy, see "Risk-based inspections as of 2009" (Inspectie van het Onderwijs 2009a, b).

1.3 Research questions

The aim of the current study has been to explore a methodology to detect which schools are at risk of underperformance for the purpose of risk-based educational accountability

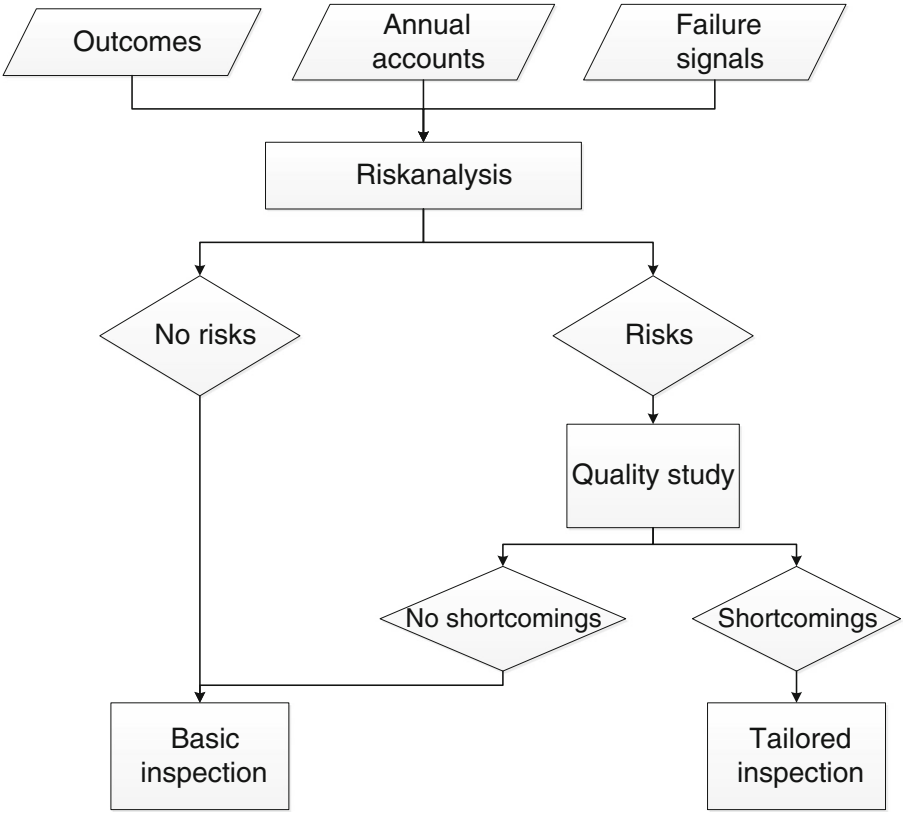

Fig. 1 Schematic overview of risk-based inspection in Dutch primary education

in Dutch primary education. Our study has therefore concentrated on finding answers to the following research questions:

1. What factors and which levels of these factors are relevant in predicting "at-risk" schools in Dutch primary education?
2. To what extent can a risk model distinguish between satisfactory performing schools and underperforming schools?

In order to find answers to these two research questions we also assessed the robustness of a risk model.

2 Method

2.1 Datasets

In this exploration of risk analysis, we tried to use as much data as possible from the Dutch Inspectorate of Education. In order to estimate which schools are "at risk," we thought it important to base the risk analysis on data that allowed us to estimate the

schools' value added. Because in a world of arguably imperfect school performance indicators, there is at least some clear justification for interpreting value-added measures as "an indication of the extent to which any given school has fostered the progress of their students in a range of subjects during a particular time period in comparison to the effects of other schools in the sample" (Sammons et al. 1997, p. 24). As student-level data on both prior and final achievement, required to investigate the schools' value added, were not (yet) available by the Dutch Inspectorate of Education, we sought alternative sources of information. This study therefore deviates somewhat from the current practice of educational accountability in the Netherlands. Still, it might provide a useful pilot study on statistical approaches to examine and identify risk factors in relation to school underperformance.

The first source of information in this empirical study concerned a student-level dataset from a sample of primary schools derived from the Monitoring and Evaluation system of CITO, the Netherlands Institute for Educational Measurement. The data are collected by schools for their own use, and these data are therefore not available for the Inspectorate of Education to use in the risk analysis. The monitoring system the data were derived from offers schools and teachers the possibility to monitor the progress of their students during primary education via several instruments, such as a set of tests, a registration system, remediation guidance methods, and tools for identifying specific learning problems. The data in our study contained reading comprehension test scores from grade 3 until grade 5 (age approximately 8–11 years) of students in Dutch primary education from 2003 to 2011. The specific focus on grade 3 to grade 5 was based on the availability of student test scores. In the early years (grades 1 and 2) and in the final year (grade 6) of primary education, only a small number of students are tested by their schools. In the early years, generally, more emphasis is given to the technical aspects of reading, while in the final year, almost all students make a different test that could not be linked to our dataset. This implies that we cannot cover the complete formal stage of schooling during primary education with our analysis. A major limitation of this data is that although the students' age and gender were recorded, other student background characteristics, such as ethnicity and socioeconomic status, were not available.

From the dataset, it was possible to construct six consecutive cohorts of students from grade 3 to grade 5. Table 1 presents the number of students and schools in each of these cohorts. For example, "Cohort 2003" consists of 7693 students that were in grade 3 in the school year 2003/2004 and that were tracked in the following two school years. It is apparent from Table 1 that not all students could be followed these 3 years. Of the 7693 students of cohort 3, 7316 students were found in school year 2004/2005 and 7020 students were found in the year 2005/2006. A similar decrease in the number of students was found for all cohorts. To ensure that our value-added and final achievement performance indicators were sufficiently reliable, we constructed three combined cohorts (Table 1), which resulted in a larger number of students per school. For example, cohort 2003 and cohort 2004 were combined into cohort 2003/2004. This combined cohort consisted of students that were in grade 3 either in 2003/2004 or in 2004/2005 and that were followed 2 years after being grade 3. The students and schools were selected for the combined cohorts based on the following criteria: (1) identification variables had to be available at both the student level and the primary school level, and (2) test scores had to be available for the schools of both individual cohorts. Students who had retained a grade were kept in the sample.

Table 1 Samples for reading comprehension

Cohort		2003/2004	2004/2005	2005/2006	2006/2007	2007/2008	2008/2009	2009/2010	2010/2011
Cohort 2003/2004: 15,195 students and 262 schools									
2003	Grade	3	4	5					
	Students	7693	7316	7020					
	Schools	265	264	257					
2004	Grade		3	4	5				
	Students		8458	7904	7514				
	Schools		299	290	288				
Cohort 2005/2006: 17,886 students and 314 schools									
2005	Grade			3	4	5			
	Students			8881	8256	7933			
	Schools			314	304	304			
2006	Grade				3	4	5		
	Students				10,061	9383	9072		
	Schools				338	334	330		
Cohort 2007/2008: 22,815 students and 371 schools									
2007	Grade					3	4	5	
	Students					11,390	10,645	8946	
	Schools					375	370	314	
2008	Grade						3	4	5
	Students						12,755	10,496	6065
	Schools						416	356	220

The sample of schools in this dataset was relatively homogeneous and not fully representative of the population of primary schools in the Netherlands. Based on the total inspection framework as described above; 3.7 % of the schools in the sample were considered as inadequate and 0.4 % of the schools as very poor. At the same time, in the population, 6 % of the primary schools in the Netherlands were considered as inadequate and 1.5 % as very poor (Inspectie van het Onderwijs 2012). Similarly, our dataset contained schools with a relatively high proportion of students from highly educated parents.

The second dataset in the empirical analysis contained school-level information regarding the characteristics of the schools until the year 2009. This dataset was derived from the Dutch Inspectorate of Education and is currently being used in the risk analysis in the Dutch educational accountability system. It consists of a rich set of variables concerning the curriculum, classroom practices, additional support, monitoring progress, general quality, staff and student population characteristics, school board, and nonmalleable school characteristics. Variables concerning the curriculum, classroom practices, additional support, and monitoring progress were derived from school inspections records. During these school inspections, inspectors visited lessons and interviewed teachers and principals, using a standardized method and framework to assess the quality of the schools. In this study, we used the main inspection results, i.e., the assessment results regarding the various aspects of educational quality, supplemented by a number of indicators of quality assurance, which are the most commonly assessed indicators. This study is based on the most recent available inspection results

per school. Not all schools are being inspected every year via the complete framework. In general, the most recent inspection results are between 1 and 4 years old, while the oldest available ones date from the year 2003. This situation implied that the possible risk factors were not available for each of the subsequent combined cohorts, as would have been in the ideal case. This school-level data stems from the period before the introduction of a risk-based strategy in the Dutch primary education accountability system. In this period, we do not expect an association among the availability of recent school inspection records and the schools' performance. This might change after the introduction of a risk-based inspection strategy.

2.2 Instruments and variables

Two types of variables were used. First, student-level variables to estimate the performance of the schools for the successive cohorts, and second, school-level variables as predictors in the risk analysis. The variables used to estimate the indicators of value added and final achievement were:

Reading comprehension We used tests from the CITO Monitoring and Evaluation system, administered by the schools to monitor the performance of their students. In grades 3, 4, and 5, the level of students' reading comprehension is generally measured by administering a grade-specific test developed by CITO. For each test, there is a paper and digital version. In these tests, the students' comprehension and interpretation of written texts are measured through 50 items. The tests contain different types of texts (e.g., informative, fiction) and different types of genres (e.g., poem, letter, story, article). The scores of the students on these grade-specific tests can be converted to a single latent one-dimensional reading comprehension scale (Feenstra et al. 2010). The reliability rates of the reading comprehension tests in the Monitoring and Evaluation systems are between 0.84 and 0.93 for the paper versions and between 0.83 and 0.93 for the digital versions.

Time The time variable for the growth models was constructed based on the exact date on which a student made a test and the end of grade 5 (1st of June). Time was expressed as the difference in years between those two dates. A value of zero indicates that the test was made at the 1st of June at the end of grade 5. The time variable has negative values for tests made before the end of grade 5. The end of grade 5 was chosen as reference in order to allow value-added and final achievement estimates to be drawn from one multilevel model (see Section 2.3).

Cohort This variable indicated whether a student in the combined cohorts belonged to the first year (for example cohort 2003 from the 2003/2004 cohort) or to the last year.

The second set of variables are the school-level variables that were derived from data from the Inspectorate of Education. Due to the long list of available risk factors, a description of all variables is included in Table 2 and descriptive statistics of these variables are presented in Tables 3 and 4. All the variables described in Table 2 are used as predictors of underperformance of schools in the risk analysis.

Table 2 Description of the variables used in the risk analysis

Name	Description
Goals	Inspectors' judgment whether all goals for Dutch language and Mathematics are covered in the curriculum. (1=sufficient; 0=insufficient)
Adaptive arrangements	Inspectors' judgment whether schools with a high proportion of students from low educated parents provide adaptive teaching arrangements for Dutch language (1= sufficient; 0=insufficient)
Clear explanation	Inspectors' judgment whether the teacher explains clearly (1=sufficient; 0= insufficient)
Student engagement	Inspectors' judgment whether students are engaged with the learning activities (1= sufficient; 0=insufficient)
Task-oriented atmosphere	Inspectors' judgment whether the schools has a task-oriented working atmosphere (1= sufficient; 0=insufficient)
Approach for extra support	Inspectors' judgment whether the school uses a systematic approach to providing extra support (1=sufficient; 0=insufficient)
Extra support evaluation	Inspectors' judgment whether the school regularly evaluates effects of extra support (1=sufficient; 0=insufficient)
System of tests	Inspectors' judgment whether the school uses a coherent system of instruments and testing for monitoring progress of students (1=sufficient; 0=insufficient)
Monitoring progress	Inspectors' judgment whether teachers monitor the progress of students systematically (1=sufficient; 0=insufficient)
Progress of development	Inspectors' judgment whether teachers monitor the progress in the development of students systematically (1=sufficient; 0=insufficient)
Student performance evaluation	Inspectors' judgment whether the school performs an annual evaluation of the performance of students (1=sufficient; 0=insufficient)
Learning process evaluation	Inspectors' judgment whether the school regularly evaluates the learning process (1= sufficient; 0=insufficient)
Overall judgment	The overall judgment is a composite variable based on the previous performance results of schools as measured by previous performance indicators and an assessment made during the school inspections (3 ordinal categories). The majority of schools were judged satisfactory (coded 0). Schools were judged as inadequate when the performance indicators or the quality of their processes were considered insufficient (coded 1). The schools were judged as very poor when both the performance indicators and the quality of their processes were deficient (coded 2)
Number staff	Indication of the size of the school as measured by the number of employees
Younger staff	Indication of the experience of the schools' staff as measured by percentage of staff under 30 years of age
Older staff	Indication of the experience of the schools' staff as measured by percentage staff over 56 years of age
Female staff	Indication of the gender distribution of staff in the school as measured by the percentage of female staff
Growth staff	Indication of growth of the school as measured by the percentage of growth in the number of employees within the past year
Intake from outside education	Indication of staff movement in relation to the school and indication of inexperienced new teachers or other staff. Measured by percentage of staff intake for outside primary education within the past year
Intake from other schools	Indication of staff movement in relation to the school. Measured by percentage of staff intake from other primary schools within the past year
Leaving outside education	Indication of staff movement in relation to the school. Measured by percentage of staff leaving outside primary education within the past year

Table 2 (continued)

Name	Description
Leaving to other schools	Indication of staff movement in relation to the school. Measured by percentage of staff leaving to other primary schools within the past year
Supporting staff	Indication of the distribution of the staff over positions as measured by the percentage of supporting staff
Part-timers	Indication of the distribution of the staff over positions as measured by the percentage of part-timers
Management	Indication of the distribution of the staff over positions as measure by the percentage of management staff
School board	A dummy variable, referring to whether the school's board had one or more schools under its supervision. In general, school boards responsible for multiple schools are usually considered to be more professional, as they have better access to support resources. Furthermore, they can cope more easily with financial, staff, and governmental issues than smaller boards responsible for only one school
Number students	Indication of the size of the school as measured by the number of students
Growth students	Indication of growth of the school as measured by the percentage of growth in the number of students within the past year
Highly educated parents	Indication of the schools' student population as measured by the percentage of students with highly educated parents
Low educated parents	Indication of the schools' student population as measured by the percentage of students with low educated parents
Very low educated parents	Indication of the schools' student population as measured by the percentage of students with very low educated parents
Dutch colonies	Indication of the schools' student population from traditional minority groups, as measured by the percentage of students from Dutch colonies (Aruba, the Netherlands Antilles, and Suriname)
Turkey and Morocco	Indication of the schools' student population from traditional minority groups, as measured by the percentage of students from Turkey or Morocco
12-year olds	Indication of the amount of grade retention in the school, measured by the percentage of 12-year-old students at the beginning of the school year
Denomination	Indication of the religious believes of the school, measured in four categories, namely public, Catholic, Protestant, and other schools
Educational vision	Indicator of philosophical and ideological vision of schools on education and a child's learning, measured in five categories: regular, Dalton, Jenaplan, Montessori, and other. Dalton, Jenaplan, Montessori schools deviate from the regular schools by making different choices in curriculum, school organization, basic activities and time management, management of space
Urbanization	Indication whether the primary schools were located within or outside large cities on the basis of the following categories: 4 largest cities of the Netherlands, 32 largest cities of the Netherlands, and outside the large cities

2.3 Method of analysis

Estimating value added and final achievement The first step in the process of our risk analysis was to estimate the performance of schools based on the test scores of reading comprehension. School performance was indexed by the achievement estimated at the end of grade 5 and the value added measured from grade 3 until grade 5. However, it

Table 3 Descriptive statistics of the variables used in the risk analysis

Variables	*N* in 2009	% of schools judged as insufficient
Goals	460	3.5
Adaptive arrangements	460	4.8
Clear explanation	462	3.7
Student engagement	461	3.5
Task-oriented atmosphere	461	2.6
System of tests	483	7.9
Approach for extra support	483	31.7
Monitoring progress	451	8.6
Progress of development	181	33.1
Extra support evaluation	470	39.8
Student performance evaluation	478	32.2
Learning process evaluation	474	32.1
Overall judgment	461	
Satisfactory		95.9
Inadequate		3.7
Very poorly		0.4

must be noted that this specific focus on grade 3 to grade 5 may have disadvantaged schools which performed better than average in these respects up to grade 3. Both performance indicators were estimated using one multilevel growth model with measurement occasions (level 1) nested within students (level 2) and students nested within schools (level 3). The models were estimated using the MLwiN 2.25 software (Rasbash et al. 2009) by modeling the scores on the reading comprehension tests as a function of time. Since multilevel growth models do not require a strictly balanced design, they can easily cope with data missing at one or more measurement occasions (Quené and Van den Bergh 2004; Snijders and Bosker 2012). The final achievement indicator could be derived from the multilevel growth model through the school-level intercept residuals, because a zero value of the time variable indicated that the test was made at the end of grade 5. The final achievement indicator therefore reflected the difference between students' performance in a particular school at the end of the fifth grade and the average performance of students in the sample at the end of grade 5. The value-added indicator could be derived from the growth models through the school-level slope residuals of the time variable. The operationalization of value added in these multilevel models was then the difference between the average progress of student performance in the sample and the progress of student performance in a particular school. Whether a school performed significantly above or below the average on final achievement and/or value added was established by testing whether the school-level residuals statistically differed from zero (average). Schools were assigned to the group of average performing schools if their value-added and their final achievement residuals were not significantly different from zero.

For the following risk analysis, schools were defined as underperforming on the indicators when they performed significantly below average on final achievement and/

Table 4 Descriptive statistics of the variables used in the risk analysis

Variables	*N* in 2009	Mean	SD	Percent
Number staff	497	22.09	10.97	
Younger staff	497	20.20	12.26	
Older staff	497	4.81	5.59	
Female staff	497	81.64	8.56	
Growth staff	497	3.99	15.01	
Intake from outside education	497	10.32	21.69	
Intake from other school	497	4.38	8.20	
Leaving outside education	497	7.92	18.30	
Leaving to other school	497	3.83	6.53	
Supporting staff	497	10.58	7.57	
Part-timers	497	9.47	5.09	
Management	497	50.13	15.31	
Number students	500	244.62	133.25	
Growth students	497	1.49	17.55	
Highly educated parents	500	0.89	0.13	
Low educated parents	500	0.06	0.06	
Very low educated parents	500	0.04	0.10	
Dutch colonies	500	3.47	9.31	
Morocco and Turkey	500	10.47	30.95	
12-year olds	500	0.02	0.01	
School board	500	0.07	0.255	
Denomination				
Public				38.0
Catholic				24.8
Protestant				27.6
Other				9.6
Educational vision				
Regular				87.7
Dalton				2.6
Jenaplan				4.0
Montessori				3.2
Other				2.4
Urbanization				
4 largest cities				18.8
32 next largest cities				8.4
Outside large cities				72.8

or value added. There are several reasons for the combination of the two performance indicators. In the first place, the judgment about the performance of schools in Dutch educational accountability is based on multiple school performance indicators (Inspectie van het Onderwijs 2011) in order to prevent strategic behavior of schools.

Combining final achievement and value added may lead to a rather heterogeneous group of schools.

Risk analysis The second part of the study was the actual risk analysis. For predicting the current underperformance of schools (t; cohort 2005/2006), we used all information available on possible risk factors (Table 2) and school performance indicators derived from the 2003/2004 cohort ($t-1$ and before). Regression tree analysis was applied to detect important interactions among possible risk factors in the schools' current performance (Neville 1999). This approach identified those risk factors that differentiated the most between underperforming schools and schools with average or above average performance. The chi-square automatic interaction detector (CHAID) algorithm used found these differences by applying χ^2 tests to measure the association between the dependent variable (adverse effect) and the independent variables (risk factors) (Agresti 1990). The CHAID algorithm does not exclude missing data. Instead, missing data are handled as a separate category, which can be combined with one or more other categories if they are statistically homogeneous. This method of risk analysis resulted in a number of risk factors associated with underperformance of schools. Furthermore, the regression tree analysis yielded a number of end nodes (relatively homogeneous subgroups of schools) based on particular combinations of the risk factors.

The extent to which this regression tree risk model distinguishes between satisfactory performing and underperforming schools can be investigated by the number of correct and incorrect classifications of schools as underperforming or satisfactory performing. In a risk analysis, the regression tree model provides each school's risk level on underperformance, which is equal to the proportion of underperforming schools in a particular end node. For this study, we considered two classification rules to determine whether schools, in particular, end nodes, were "at risk." The first classification rule was a probability of underperformance higher than 0.50 (over half of the schools in an end node underperformed on final achievement and/or value added). In regression tree analysis, 0.50 is the standard rule. Given this classification rule, all schools in an end node with 50 % or more underperforming schools were considered "at risk," and all schools in an end node with a probability of lower than 0.50 were regarded "not at risk." The second classification rule was a probability of underperformance higher than 0.10, indicating that over a tenth of the schools in an end node underperformed. A rule based on a probability of 0.10 could be considered more conservative than the 0.50 rule. When applying this rule, all schools in an end node with a probability of 0.10 and higher were regarded to be at risk and those with a probability of lower than 0.10 not at risk. Given the previous classification rules the results of the regression tree analysis can be analyzed in terms of *false positives* and *false negatives*. In our study, a false positive was a school in an end node which was considered "at risk" as indicated by the risk factors, whereas the observed performance of this school was not significantly below average. False negatives were those schools that underperformed in 2005/2006 (at time t), whereas the model did not predict any potential risk as based on the risk factors (until $t-1$). These were underperforming schools in an end node where there was relatively little underperformance. The

fewer false positives and false negative the regression tree model yields, the better it is possible to distinguish between satisfactory performing schools and underperforming schools.

To test whether the risk model was robust over time, the regression tree rules from the analysis of cohort 2005/2006 were applied to the performance data of the 2007/2008 cohort. Again, the results were rendered in terms of false positives and false negatives. Increasing numbers of false positives and/or false negatives would imply that the model was not robust over time. In this way, it could be determined whether or not the risk model may be useful in the practice of educational accountability.

3 Results

3.1 Differences in final achievement and value added among Dutch primary schools

The results of the multilevel growth models for the estimation of value added for the three subsequent combined cohorts for reading comprehension are presented in Table 5. For final achievement at the end of grade 5, between 9.2 % (cohort 2005/2006) and 11.5 % (cohort 2003/2004) of the total variance was accounted for by the school level. These differences in intercepts among schools at the end of grade 5 indicate that the schools differed in terms of their final achievement in reading comprehension. The between-school differences appeared to be somewhat smaller than those found in previous research in similar grades in the Netherlands (i.e., 15 % in grade 4 in Bosker et al. 1997; 17.6 % in grade 5 in Verhelst et al. 2003; 14.2 % in grade 6 in Wijnstra et al. 2003), which may be due to the relatively homogeneous sample in our analysis. The slope differences for the time variable indicate that the schools also differed in their value added for reading comprehension. This between-school variance in slopes was the largest in the 2007/2008 cohort. For 95 % of the schools in the 2007/2008 cohort,[1] the progress in reading comprehension ranged between 4.87 and 13.44 points on the latent reading comprehension scale. This finding implies that in a school with a high value added, the progress of the students is over 2.5 times that of the progress of students in schools with a low value added.

When looking at between primary school differences at the end of grade 5 (final achievement) and the progress over time (value added), positive associations were found for all combined cohorts (for example $r=0.57$; $N=371$; $p<0.001$; cohort 2007/2008). These results imply that schools with a high final achievement also tend to show more progress over time. When looking at primary school differences at the start of grade 3 and the progress over time (value added), negative correlations were found for all three subsequent cohorts (for example $r=-0.65$; $N=371$; $p<0.001$; cohort 2007/2008). These negative correlations imply that student performance tends to grow

[1] School differences in the progress of reading comprehension were calculated using the following formula: $9.15 \pm 1.96\sqrt{4.87}$ (Snijders and Bosker 2012, p. 53). In this formula, 9.15 indicated the average progress of the students in reading comprehension in cohort 2007/2008, and 4.87 the between-school variance for the slope of time.

Table 5 Multilevel growth models for estimating value added of primary schools for reading comprehension

	Cohort 2003/2004		Cohort 2005/2006		Cohort 2007/2008	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Fixed part						
Intercept	51.358	0.382	51.955	0.314	47.741	0.309
Time in years	9.885	0.134	9.907	0.122	9.152	0.130
Cohort	0.669	0.203	0.873	0.186	−0.274	0.162
Random part						
School-level intercept variance	29.779	3.084	22.241	2.219	25.611	2.402
School-level slope variance for time	3.764	0.405	3.695	0.366	4.872	0.450
School-level covariance intercept and slope	6.351	0.921	4.126	0.706	6.767	0.867
Student-level intercept variance	162.103	3.063	155.934	2.734	153.229	2.830
Student-level slope variance for time	4.752	0.603	3.950	0.535	2.600	0.576
Student-level covariance intercept and slope	9.486	1.071	8.337	0.950	9.736	1.033
Measurement occasion variance	66.617	0.784	64.606	0.705	69.124	0.725
Model fit and sample size						
−2*log-likelihood	339,340.025		394,705.054		450,852.407	
Number of schools	262		314		371	
Number of students	15,195		17,886		22,815	
Number of measurements	43,582		50,921		57,863	

less in schools with a high initial achievement in grade 3. The latter results may suggest a ceiling effect in the test scoring. Overall, the results of the multilevel growth models indicate a convergent pattern, with more variability in performance among schools in grade 3 compared to grade 5. This convergent pattern is similar to previous research on between-school differences in comparable grades in Dutch primary education (Guldemond and Bosker 2009).

3.2 Risk analysis based on regression tree analysis

Table 6 and Fig. 2 present the results of the regression tree analyses for investigating which set of characteristics were related to underperformance in reading comprehension for the 2005/2006 cohort. Of the total sample of primary schools of the 2005/2006 cohort, 76 schools (24.2 %) were identified as underperforming on final achievement and/or value added. This percentage of 24 % underperforming schools is much higher than the 4.1 % of the schools with an overall inadequate or very poorly judgment from the Inspectorate of Education in 2009 (Table 3). Several differences in the definition and operationalization of underperformance may have added to this dissimilarity, such as the used data and constructs (student- vs. school-level data; performance indicators vs. performance and process indicators), construction of performance indicators (two vs. five performance indicators; achievement and progress vs. achievement and efficiency), and rules to combine several indicators into a composite variable (and/or rule vs. and rule).

Table 6 Results of the regression tree for predicting underperformance as regards reading comprehension in cohort 2005/2006

Node	Underperforming schools		Parent node	Primary independent variable				
	N	Percent		Variable	Sig.	Chi-square	*df*	Split values
0	76	24.2						
1	20	64.5	0	Proportion of students from highly educated parents	0.000	39.457	2	≤0.74
2	29	30.5	0		0.000	39.457	2	[0.74–0.90]
3	27	14.4	0		0.000	39.457	2	>0.90
4	16	16.3	3	Final achievement 2003/2004	0.000	18.894	2	Average
5	6	54.5	3		0.000	18.894	2	Underperforming
6	5	6.3	3		0.000	18.894	2	Overperforming and missing
7	11	29.7	4	Regular evaluation effects of extra support	0.016	7.817	1	Insufficient
8	5	8.2	4		0.016	7.817	1	Sufficient and missing
9	6	60.0	7	Annual evaluation of the performance of students	0.014	6.010	1	Insufficient
10	5	18.5	7		0.014	6.010	1	Sufficient
11	3	30.0	8	The use of a systematic approach of providing extra support	0.018	7.556	1	Insufficient
12	2	3.9	8		0.018	7.556	1	Sufficient and missing

This table provides the details of the regression tree analysis that is presented in Fig. 2

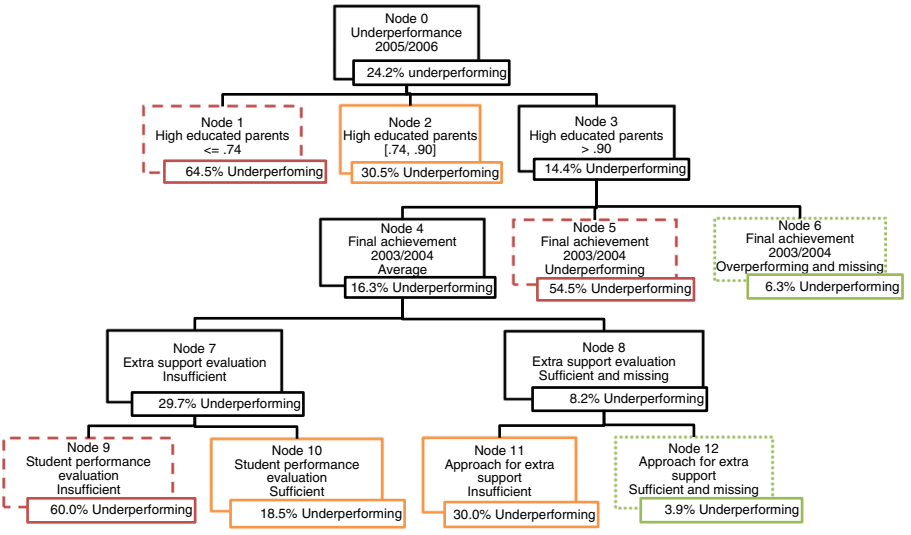

Fig. 2 Results from the regression tree analysis

The total 2005/2006 sample can be divided into more homogeneous subsamples based on five splitting variables, which resulted in eight end nodes. These end nodes of the regression tree in Fig. 1 are colored and dashed based on the relative number of underperforming schools in each node. The green-colored end nodes (dotted lines) contain a subgroup of schools that consists of relatively few underperforming schools (i.e., less than 10 % of the schools are underperforming). There are two green-colored dotted-lined end nodes. The first green end node is 6, which consists of schools with relatively many students from highly educated parents and with high or missing past performance records (6.3 % underperforming schools). The second green end node is 12, which consists of schools with relatively many students from highly educated parents, average final achievement in 2003/2004, sufficient judgments of the evaluation of extra support, and sufficient judgments on having a systematic approach in offering extra support to students (3.9 % underperforming schools). Together, these two end nodes contain 130 primary schools. A basic inspection regime might be given to the schools in these end nodes because of the low-risk levels.

The orange-colored end nodes contain a subgroup of schools with some underperforming schools (i.e., more than 10 % but less than 50 % underperforming schools). There appear three orange end nodes, namely end nodes 2, 10, and 11. End node 2 consists of schools that have a student population in the range of more than 74 % students from highly educated parents until 90 % student from highly educated parents (30.5 % underperforming). Descriptive statistics showed that the proportion of students from highly educated parents was mainly related to the final achievement indicator (r=0.46; N=314; p<0.001) and not to the schools' value added (r=−0.01; N=314; p<0.842). The second orange end node is node 10, which consists of schools with over 90 % students from highly educated parents, average final achievement in 2003/2004, insufficient judgments of the evaluation of extra support, and sufficient judgments on having a systematic approach in offering extra support to students (18.5 % underperforming schools). The last orange end node (11) consists of schools with relatively many students from highly educated parents, average final achievement in 2003/2004, sufficient judgments of the evaluation of extra support, and insufficient judgments on having a systematic approach in offering extra support to students (30.0 % underperforming schools).

Finally, the red-colored end nodes (dashed lines) contain a subgroup of schools that consists of many underperforming schools (i.e., over 50 % of the schools are underperforming). Three end nodes from the regression tree analyses contain over 50 % underperforming schools. The schools in these three red end nodes could be regarded as a high-risk schools that may receive a more intensive inspection regime. The first red end node is node 1, which consists of schools with relatively few students from highly educated parents (64.5 % underperforming schools). The second red end node is node 5, which contains schools with relatively many students from highly educated parents and with low past performance records (54.5 % underperforming). The third red end node is node 9, which contains schools with relatively many students from highly educated parents, with average past performance records, insufficient judgments on the evaluation of extra support, and also insufficient judgment on the evaluation of their students' performance (60.0 % underperforming).

Table 7 presents the classification of schools as "at risk" and "not at risk" as predicted by the regression tree, based on all known information until the 2003/2004

Table 7 Classifications in cohort 2005/2006

		Cohort 2005/2006	
		Observed underperformance	
		No	Yes
Estimated risks 0.50 classification rule	Not at risk	218 (69.4 %) True negatives	44 (14.0 %) False negatives
	At risk	20 (6.4 %) False positives	32 (10.2 %) True positives
Estimated risks 0.10 classification rule	Not at risk	123 (39.2 %) True negatives	7 (2.2 %) False negatives
	At risk	115 (36.5 %) False positives	69 (22.0 %) True positives

cohort, in comparison with their observed performance in 2005/2006. The end nodes 1, 5, and 9 consisted of over 50 % of underperforming schools (red), and schools in these end nodes were considered "at risk" under the 0.50 classification rule. These three end nodes included in total 52 primary schools, which was 17 % of the total sample. Less than half of the observed underperforming schools in 2005/2006 were found in these three end nodes (*true positives*; 32 schools; 42.1 %). Under the 0.50 classification rule, more than half of the observed underperforming schools were not considered as "at risk" based on the regression tree analysis (*false negatives*; 44 schools; 57.9 %).

The end nodes 1, 2, 5, 9, 10, and 11 (red and orange) contained over 10 % of underperforming schools which could be regarded as "at risk" under the 0.10 classification rule. These six end nodes included in total 184 schools, which was 59 % of the total sample. The nodes contained 69 of the 76 primary schools which underperformed in reading comprehension for the cohort 2005/2006 (*true positives*; 90.7 %). Seven underperforming primary schools in 2005/2006 were, however, not found in these nodes (*false negatives*; 9.2 %). Furthermore, 115 primary schools in these end nodes, and thus considered "at risk," did not underperform in the 2005/2006 cohort (*false positives*). The fact that for reading comprehension six end nodes from the regression analysis contained more than 10 % underperforming schools (red and orange) implies that as regards this subject there is no single set of characteristics which adequately identifies underperforming schools.

3.3 Applying the risk model to the performance data of the 2007/2008 cohort

The robustness of the risk model was investigated by applying the splitting rules of the regression tree analysis based on the 2005/2006 cohort to the performance data of the schools of cohort 2007/2008. Table 8 shows the results of this robustness check in terms of false and true positives and negatives. For this robustness check, we applied the 0.50 and 0.10 classification rules again. By applying the 0.50 classification rule to the performance data of cohort 2007/2008 (red end nodes 1, 5, and 9), we found that 18.1 % of the schools were considered "at risk." This percentage was similar to the previous 2005/2006 cohort (17 %). In these three end nodes, 24 underperforming

Table 8 Classifications in cohort 2007/2008

		Cohort 2007/2008	
		Observed underperformance	
		No	Yes
Estimated risks 0.50 classification rule	Not at risk	252 (67.9 %) True negatives	52 (14.0 %) False negatives
	At risk	43 (11.6 %) False positives	24 (6.5 %) True positives
Estimated risks 0.10 classification rule	Not at risk	139 (37.5 %) True negatives	14 (3.8 %) False negatives
	At risk	156 (42.0 %) False positives	62 (16.7 %) True positives

schools were found, which was 31 % of all underperforming schools in this cohort. Compared to the 2005/2006 cohort, even more of the observed underperforming schools were not considered as "at risk" under the 0.50 classification rule (*false negatives*; 52 schools; 68.4 %).

Using the 0.10 classification rule, we observed that 218 schools of cohort 2007/2008 were eligible for further investigation, which was 59 % of the total sample. This finding showed that applying the regression tree model to the data of cohort 2007/2008 did not lead to an increase in the relative amount of schools which could be considered as schools at risk. Based on these rules, 81.6 % (62 of the 76) of the underperforming primary schools in the 2007/2008 cohort were found in the six end nodes that were considered "at risk" (red and orange). Furthermore, there were 14 false negatives: schools with observed underperformance in 2007/2008 but that were not found in the six end nodes and thus not considered "at risk." Applying the model to the data of a subsequent cohort resulted in an increase in the number of false negatives: 7 in cohort 2005/2006 and 14 in cohort 2007/2008. Furthermore, the model resulted in 156 schools with no observed underperformance but considered "at risk" in the 2007/2008 cohort (*false positives*), which was 42.0 % of all primary schools.

4 Conclusion and discussion

The aim of the current study has been to explore a methodology to detect which schools are at risk of underperformance for the purpose of risk-based educational accountability in Dutch primary education. Within the context of educational accountability, low student performance at the end of a formal stage of education at schools as well as a limited progress during this trajectory at schools were investigated as adverse effects. In this study, we therefore defined a school "at risk" as a school which had based on their characteristics and previous performance risk factor a high change of having low final academic achievement and/or a low value added. Regression tree analysis was applied to predict which schools were "at risk," using a rich set of school level characteristics (risk factors), for

example student and staff composition, previous performance, and results of school inspections.

Some European inspectorates of education have shifted toward a risk-based school inspection system, because this approach has been assumed to be more efficient and effective than the traditional inspection strategies (Sparrow 2000). This study's results indicate that if the regression tree risk model was going to be applied in the context of educational accountability in Dutch primary education, additional information and a quality study would be requested for 59 % of the primary schools in cohort 2005/2006. However, in our study, a group of about 40 % of the schools was identified which showed very small risks. These results represent the moderate efficiency gain yielded when risk-based school accountability models are applied in Dutch primary education. When we applied the rules from the regression tree analysis to the performance data of schools in a later cohort (the robustness check), the number of false negatives increased. This means that an increasing number of schools which underperformed in 2007/2008 slipped through the nets of the analysis, as they were not classified as schools "at risk."

All in all, the results of this risk analysis therefore imply that the underperformance of schools cannot be predicted very accurately. This relates to the issue of the moderate stability of school performance indicators over subsequent cohorts (e.g., Leckie and Goldstein 2009; Mortimore and Sammons 1994; Thomas et al. 2007; Van der Werf and Guldemond 1996; Wilson and Piebalga 2008). The moderate stability of school performance indicators attenuate the extent that a risk analysis can identify underperforming schools and thereby the possible efficiency gain. Although differences in school performance indicators across time might indicate actual changes in the performance of a school, they might also reflect unreliability of the estimation of value added. Unfortunately, there are no easy solutions to this issue. One direction to go may be to conduct further research into the reliability and validity of school performance indicators (e.g., Gorard 2006, 2008; Kelly and Downey 2010; Martineau 2006). A second direction may be to combine subsequent cohorts of students to improve reliability, which is the usual method in Dutch educational accountability and has also been applied in this study. Although one might hope for more accurate classifications of underperforming schools, the results of the risk models based on final achievement and valued added provide some information to realize a more efficient educational accountability strategy.

The second finding of this study is that the number of false positives and false negatives in a risk analysis depends heavily on the decision rules that determine which schools are considered "at risk." Schools were considered at risk if their probability on underperformance, as based on their known characteristics, was higher than 0.50 (high risk) or 0.10 (some risks). When applying the 0.10 rule, the number of false negatives appeared to be relatively small while that of false positives became relatively large. Based on these rules, almost all underperforming schools were located, and therefore, we might expect an effectiveness gain since the schools that benefit most can be targeted; however, this is at the expense of the efficiency gain. Through applying less conservative rules (0.50 rule), fewer schools are considered "at risk," resulting in an increase in the number of false negatives. Therefore, the less conservative rules seem efficient, but this less conservative rule attenuates the possible effectiveness gain because a relatively large part of the schools that might benefit the most from school inspections slip through the nets. The current policy of the Dutch Inspectorate of

Education is that only a very small number of false negatives can be allowed in order to prevent underperforming schools to slip through the nets, and therefore, a more conservative classification rule (0.10 rule) needs to be applied.

The third finding in this study is that underperformance of primary schools could be explained by the following risk factors: the composition of the school in terms of the proportion of students from highly educated parents, previous final achievement, and the inspectors' judgments on evaluation of the effects of extra support, monitoring of student performance, and the use of a systematic approach to the provision of extra support. An important finding is that three sources of information were necessary to determine schools at risk, namely the schools' previous performance, annual accounts on staff and student populations, and the quality of the schools' processes as judged by the inspectors. The moderate efficiency gain found in the regression tree could only be reached if an Inspectorate of Education records and analyzes all these three types of data of primary schools systematically. This is another indication that good and reliable data are crucial for risk-based inspection strategies (Hulett and Preston 2000).

Finally, the results of the regression tree analysis, showing multiple end nodes with high risks, further indicated that "at-risk" schools cannot be described by a uniform set of characteristics. There were three high-risk end nodes (red) and three some risk end nodes (orange), each predicted by a different set of risk factors. This finding implies that the statistical model in a risk-based educational accountability needs to include the possible interactions among risk factors. Regression tree analysis or other statistical models that cluster schools into relatively homogeneous subgroups and in which interactions among risk factors can be explicitly modeled seem therefore appropriate models. However, these statistical models are highly complex, and therefore, they may become a black box for many stakeholders in education.

When interpreting the results, several limitations of the study need to be kept in mind that mostly relate to the student-level dataset that was used. By using a student-level dataset from the CITO Monitoring and Evaluation system, we deviated from the current practice in Dutch educational accountability, in order to include value added as an adverse effect in the risk model. By estimating the schools' value added, it was possible to take the students prior achievement into account and, therefore, to make a comparison between schools that was probably more fair than raw performance indicators. The sample of schools in this dataset, however, was relatively homogeneous and not fully representative of the population of primary schools in the Netherlands. Despite this homogeneous sample, we still found considerable differences among the schools in terms of their final achievement and value added. Although these between-school differences were somewhat smaller than in previous studies (Bosker et al. 1997; Verhelst et al. 2003; Wijnstra et al. 2003), the pattern of final achievement and value added was comparable to previous results in similar grades of Dutch primary education (Guldemond and Bosker 2009). Moreover, the sample of primary schools in this study was relatively small (n=371), which fairly undermined the power of the regression tree analysis, given the large number of possible predictors of underperformance.

Another major limitation of the student-level dataset is that in the estimation of value added, only the students' test scores could be used, because variables of their ethnic and socioeconomic background were not available in the student-level

dataset. In order to isolate the effect of a school on student progress adequately, however, prior achievement and family background are generally also required (Timmermans et al. 2011; Willms 1992). Because the students' family background was not included in the estimation of value added, we cannot be sure whether the value added as identified in our analysis can be solely attributed to the schools or was the result of a wider social context, and this may lead to bias in the estimation of value added. However, at the current time, this was the only dataset available in which prior and final achievement were available and that allowed for value-added and final achievement performance indicators to be estimated for several subsequent cohorts.

Furthermore, it should be kept in mind that the student-level data used in this study were administered by schools for their own use to monitor the performance of students. It is therefore possible that the conditions under which these tests are administered differed among schools as well as among tests within the individual schools. In addition, schools might differ in their selection of students included in the tests and the number of tests administered each year. If some schools test their weakest students more often than the average or above average students, the school performance indicators again might become biased.

Finally, value-added models usually estimate the progress of students during several grades, the earliest of which is often grade 3 (Linn 2007). So, value-added models control for differences in prior achievement dating back to the earliest grade included in the analysis. However, they cannot rule out the possible influence of achievement differences which occurred in kindergarten and/or grades 1 and 2. This was an issue in our research as well, since the first available test scores were derived from grade 3. Findings from previous research in Dutch primary education revealed that schools that do a great job up to grade 3 are not necessarily the same schools as schools that do a great job from grade 3 onward (Guldemond and Bosker 2009). By combining final achievement and value added in this study, we tried to overcome some of the disadvantage of schools which performed better than average up to grade 3, although the specific focus on grade 3 to grade 5 could have important implications for the generalizability of the study to the complete time span of primary education especially regarding the early years.

For risk models to become commonly used in practice, the results of risk analyses have to be consistent among multiple studies (Gibb 1997). Previous studies within the tradition of educational effectiveness research, however, have generally merely focused on the characteristics of effective schools and not on those of underperforming or failing organizations (Reynolds and Teddlie 1999). Although some recent reviews have shown consistency in school characteristics in terms of student performance, there are still inconsistencies with respect to the magnitude of the effects of these characteristics on student performance (Scheerens 2014). Based on this previous research on characteristics that relate to high performance of schools, one might expect that the results of individual studies might differ and that multiple studies into underperforming schools are needed to come with a set or sets of characteristics that relate consistently with underperformance of schools. This indicates that although this study demonstrates that a moderate efficiency gain may be achieved, risk models cannot yet be adequately applied in the practice of school accountability.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Betebenner, D. W. (2007). *Estimation of student growth percentiles for the Colorado Student Assessment Program*. Dover: National Centre for the Improvement of Educational Assessment (NCIEA).

Betebenner, D. W. (2009). *Growth, standards and accountability*. Denver: Colorado Department of Education: The Centre for Assessment.

Bosker, R. J., Lam, J. F., Dekkers, H., & Vierke, H. (1997). *De betekenis van kwaliteits verschillen tussen basisscholen*. Enschede: Universiteit Twente.

De Jong, P. (2012). The health impact of mandatory bicycle helmet laws. *Risk Analysis, 32*, 782–790.

De Wolf, I., & Verkroost, J. J. H. (2011). Evaluatie van de theorie en praktijk van het nieuwe onderwijstoezicht. *Tijdschrift voor Toezicht, 2*, 7–24.

Feenstra, H., Kamphuis, F., Kleintjes, F., & Krom, R. (2010). *Wetenschappelijke verantwoording Begrijpend lezen voor groep 3 tot en met 6*. Arnhem: CITO.

Figlio, D., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbooks in economics* (Vol. 3, pp. 383–421). The Netherlands: North-Holland.

Gibb, H. J. (1997). Epidemiology and cancer risk assessment. In V. Molak (Ed.), *Fundamentals of risk analysis and risk assessment*. Boca Raton: Lewis.

Gorard, S. (2006). Value-added is of little value. *Journal of Education Policy, 21*, 235–243. doi:10.1080/02680930500500435.

Gorard, S. (2008). The value-added of primary schools: what is it really measuring? *Educational Review, 60*, 179–185. doi:10.1080/00131910801934185.

Guldemond, H., & Bosker, R. J. (2009). School effects on students' progress—a dynamic perspective. *School Effectiveness and School Improvement, 20*, 255–268. doi:10.1080/09243450902883938.

Hulett, D. T. & Preston, J. Y. (2000). Garbage in, garbage out? Collect better data for your risk assessment. In Proceedings of the Project Management Institute Annual Seminars & Symposium (pp. 983–989). Houston.

Inspectie Onderwijs van het. (2009a). *Toezichtkader PO/VO 2009*. Utrecht: Inspectie van het Onderwijs.

Inspectie Onderwijs van het. (2009b). *Risk-based inspection as of 2009*. Utrecht: Inspectie van het Onderwijs.

Inspectie Onderwijs van het. (2011). *Analyse en waardering van opbrengsten primair onderwijs*. Utrecht: Inspectie van het Onderwijs.

Inspectie Onderwijs van het. (2012). *De staat van het onderwijs. Onderwijsverslag 2010/2011*. Utrecht: Inspectie van het Onderwijs.

Keeney, R. L., & von Winterfeldt, D. (2011). A value model for evaluating Homeland Security decisions. *Risk Analysis, 31*, 1470–1487.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: looking inside the 'black box' of complex metrics. *Educational Assessment, Evaluation and Accountability, 22*, 181–198.

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society, Series A, 172*, 835–851.

Linn, R. L. (2007). Validity of inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education*. doi:10.1007/s11092-007-9027-6.

Martineau, J. A. (2006). Distorting value added: the use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*, 35–62.

Ministry of Education. (2014). *Key-figures 2009–2013: Education, culture, science*. Den Haag: Ministry of Education.

Molak, V. (1997). *Fundamentals of risk analysis and risk management*. Boca Raton: Lewis.

Mortimore, P., & Sammons, P. (1994). School effectiveness and value added measures. Assessment in Education: Principles, Policy & Practice, 1.

Neville, P. G. (1999). Decision trees for predictive modeling. SAS Institute.

Ofsted. (2010). *The evaluation schedule for schools*. Manchester: The Office of Standards in Education.

Ofsted. (2011). *The framework for school inspection*. Manchester: The Office for Standards in Education.

Quené, H., & Van den Bergh, H. (2004). On multi-level modelling of data from repeated measures designs: a tutorial. *Speech Communication, 43*, 103–121.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A user's guide to MLwiN*. Bristol: Centre for Multilevel Modelling, University of Bristol.

Ray, A. (2006). School value added measures in England: a paper for the OECD project on the development of value-added models in education systems. Department of Education Skills.

Reynolds, D., & Teddlie, C. (1999). The future agenda of studies into the effectiveness of schools. In R. J. Bosker, B. Creemers, & S. Stringfield (Eds.), *Enhancing educational excellence, equity, and efficiency: Evidence from evaluations of systems and schools in change* (pp. 223–251). Dordrecht: Kluwer Academic.

Sammons, P., Thomas, S., & Mortimore, P. (1997). *Forging links: Effective schools and effective departments*. London: Paul Chapman.

Sanders, W. L. (2003). Beyond no child left behind. In 2003 Annual Meeting American Educational Research Association.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311.

Scheerens, J. (2014). School, teaching and system effectiveness: some comments on three state of the art reviews. *School effectiveness and School Improvement, 25*, 282–290.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles: Sage.

Sparrow, M. K. (2000). *The regulatory craft; controlling risks, solving problems and managing compliance*. Washington: Brookings Institution Press.

Standards Association of Australia. (1999). *Risk analysis*. Strathfield: Standards Association of Australia.

Thomas, S., Peng, W. J., & Gray, J. (2007). Modelling patterns of improvement over time: value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education, 33*, 261–295.

Timmermans, A. C., Doolaard, S., & De Wolf, I. (2011). Conceptual and empirical differences among various value added models for accountability. *School Effectiveness and School Improvement, 22*, 393–413. doi: 10.1080/09243453.2011.590704.

Van der Werf, M. P. C., & Guldemond, H. (1996). *Omvang, stabiliteit en consistentie van schooleffecten in het basisonderwijs*. Groningen: GION.

Verhelst, N., Staphorsius, G., & Kleintjes, F. (2003). *Scholen langs de meetlat*. Arnhem: CITO.

Wijnstra, J., Ouwens, M., & Béguin, A. (2003). *De toegevoegde waarde van de basisschool*. Arnhem: CITO.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington: Falmer.

Wilson, D., & Piebalga, A. (2008). Performance measures, ranking and parental choice: an analysis of the English school league tables. *International Public Management Journal, 11*, 344–366.