

Effectiveness of strong accountability systems

M.C.M. Ehren¹

Abstract

Most countries have some kind of education accountability system. Systems vary in indicators, standards and types of evaluation methods used, the strength of sanctions and rewards, the use of public reporting of school performance or not, and for how long accountability has been in place. The strongest accountability systems are usually expected to have a positive influence on student outcomes.

This paper describes the impact of different components and types of accountability systems on student performance. Several European countries and the U.S. were included in the study to enable comparison of different types of accountability systems, including both systems of test-based accountability and school inspections. Accountability effects on school performance were identified by relating different components and types of accountability systems to changes in TIMSS and NAEP grade-levels between 2000 and 2007. The results of the study show that accountability systems aiming at schools (instead of students, teachers or districts) have a positive impact on gains in student achievement. Some results also indicate that strong accountability systems (including a large number of standards, measurement methods, targets and stakes) and the number of measurement methods and stakes in accountability systems have a positive effect on output levels of student achievement.

Introduction

The main responsibility of educational accountability systems (both in Europe and the United States) is considered to be the monitoring of the quality of performance (compared to some standards or regulations), and ensuring that schools introduce improvements where necessary. Accountability is expected to spur teachers and schools to work harder, to work more effectively, or to reallocate instructional time and other resources from less important activities to more important ones (Koretz, 2002, p.767).

A number of comparative studies (Eurydice, 2004; CPRE, 2001) in Europe and the United States show that countries and states choose different forms and types of educational accountability to achieve this aim. Most European countries use some form of external evaluation in which evaluators (for example, Inspectorates of Education) gather information on the functioning of a school via interviews, document study and inspection visits to schools. Evaluators analyze the findings compared with national objectives and/or relative to the performance of other schools. The focus of these evaluations is mainly on *educational processes* such as classroom teaching, guidance and support of students.

The United States are an example of another type of evaluation: *output or test-based* forms of accountability. In this case the results of achievement tests are the main source of information for evaluating school performance (Koretz, 2002). Schools or educators receive rewards, sanctions or both on the basis of students' test scores. Other output parameters may be the magnitude of student absenteeism, student drop-out rates etc.

These different types of evaluations and stakes in accountability systems may also have different effects on student performance. Available research on effectiveness of accountability systems shows different results. A first strand of literature on high stakes test-based accountability systems indicates that students perform better when accountability policies are in place that aim at students (external exit exams) and schools (assessment-based comparisons). Carnoy and Loeb (2002) found for example that state external pressure on schools to improve student achievement according to state-defined performance criteria, state-wide testing in several grades, sanctions and rewards that pressure schools to improve student performance lead to an increase in achievement gains. Hanushek and Raymond (2002) compared states without any form of accountability system, states having report card systems displaying test performance and other factors and states with full accountability systems (including judgments of performance and sanctions and rewards). They found that a typical student in a state without any formal system would see a 0.7 percent increase in proficiency scores. Reporting systems

¹ University of Twente, Faculty of Behavioral Sciences, Department of Educational Organization and Management, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: m.c.m.ehren@utwente.nl

move the expected gain to 1.2 percent. Finally, states with full accountability systems obtained a 1.6 percent increase in mathematics performance.

Other authors have however also pointed to side effects of these types of accountability systems. Stecher (2001) and Koretz (2003) for example describe how teachers raise test scores by using undesirable test preparation such as reallocating classroom time to emphasize topics covered by the test or by coaching students to perform better on tests by focusing instruction on incidental aspects of the test (such as item format), or spending large amounts of class time on test-taking instruction instead of regular teaching. These types of test preparation are undesirable when they lead to increases in test scores that do not represent meaningful increases in students' knowledge of a certain domain. Other examples of side effects of test-based accountability include fraud during test administration when teachers prompt students with the right answers or when they correct wrong answers, and educational triage. Educational triage is used to increase the number of students that pass a cut-point set in the accountability system (Booher-Jennings, 2005; Gribben et al, 2008). Teachers and schools target their resources to students that are considered to be 'suitable cases for treatment' and that may improve the school's overall score. Students at the far ends of the spectrum are ignored.

Other types of accountability systems, such as school inspections, also affect education and performance of students. Studies into effects and side effects of school inspections however have shown some conflicting results with respect to types of effects. Luginbuhl et al (2007) conclude for example that test scores improve two to three percent of a standard deviation of the test score in the two years following an inspection. In line with these results Ehren (2006) found that feedback of school inspectors to schools and a negative assessment of a school led to improvement processes in the case study schools she studied. Shaw et al (2003) and Rosenthal (2004) on the other hand found a decline in student achievement levels in the year of the inspection visit. Several other authors also point to side effects of school inspections. De Wolf and Janssens (2005), Smith (1995) and Chapman (2001) describe for example how schools try to influence the assessment by implementing procedures and protocols that have no effect on primary processes in the school, but are implemented to be assessed more positively (window dressing). School inspections may also lead to a one-sided emphasis on the (quantitative) elements that are assessed, ignoring long-term goals or unquantifiable aspects of performance. Schools for example focus on programming a large number of lesson hours instead of trying to improve the quality of lessons offered.

This short overview, although not encompassing, shows that studies into effectiveness of accountability systems have produced fragmented and to some extent conflicting results. Most studies also focus on certain types of accountability systems (for example high stakes test-based accountability systems or school inspections), whereas more and more accountability systems include combinations of different types of systems. The Inspectorate of Education in the Netherlands for example uses test scores as part of a risk analysis to identify schools with low performance. These schools are scheduled for inspection visits in which their educational processes are measured and assessed.

This study contributes to the existing research by investigating in more detail which of the components of accountability systems (for example, specific types of standards, measurement methods or stakes) contribute to student achievement, and by including a wide range of accountability systems, such as test-based accountability and school inspections.

The following research questions will be addressed:

- Do stronger approaches to accountability lead to better or poorer student outcomes?
- Which components and types of accountability systems (for example, types of standards, or types of stakes) are responsible for these changes in student outcomes?

Components of accountability systems

In this section the components of accountability systems that may lead to improved student achievement are described. Several authors propose a basic skeleton of (external) accountability systems that can be used to describe these components. Hanushek and Raymond (2001), Anderson (2005) and Berry et al (2003) include goals, standards for performance, measurement, and consequences for varying levels of performance. According to Berry et al (2003) a school accountability system should contain information about an organization's performance (such as test

scores, graduation rates), standards for assessing the organization's performance, consequences for success or failure (rewards and sanctions), and an agency that collects information, judges whether or not standards have been met, and distributes rewards and sanctions. Accountability systems are, according to CCSSO (2004), used to achieve specific educational goals by attaching certain consequences to performance indicators and are meant to affect change in specific areas of functioning. The relevant components that recur in these descriptions are: stakeholders of accountability systems, measurement methods used to measure performance of stakeholders, performance standards, targets that should be met and stakes. These components will be described below.

Stakeholders

Stakeholders in accountability systems are the people in the school whose behavior is assessed and who are rewarded or sanctioned when they fail to meet the targets set. Stakeholders may be situated at different levels in the education system such as individual pupil level, class-level (including teachers), school-level (including principals and management), and system level (including policy makers). Stakeholders at student level (individual students) are for example evaluated on (nationally defined) indicators describing (minimum) performance targets each (individual) student should meet. These targets mostly include minimum test scores with respect to math, reading and writing. Students may be confronted with sanctions such as having to repeat a class or having to follow compulsory summer classes. Rewards may include scholarships for further education.

Teacher level accountability may include indicators on classroom level such as how teachers actively involve students in lessons and indicators on output level such as performance of their students on tests. Teachers may be assessed individually by school inspectors to determine whether they perform according to the indicators and performance targets in the accountability systems. In some cases, test scores of groups of students are used to assess performance of teachers. They may receive bonuses when meeting the targets, or be replaced or withheld from permanent contracts when failing to meet the targets.

Stakeholders in accountability systems are generally situated at the school-level. Indicators describe output of schools and average performance targets for the entire pupil population in a school. Indicators may also describe adequate educational processes in and organization of the school. Stakes at school level are financial rewards, compulsory assistance in school improvement, intensified monitoring, financial rewards for good performance, or closure of schools in case of long term underperformance.

The highest level in accountability systems is the system level of (sub)national administrators. Stakeholders may be districts or regions in a country. Indicators may contain the financial means, regulations, nationally implemented reform programs or educational quality of all schools in a certain region or country. System level accountability can be found in the U.S. where some states hold their districts accountable for the average performance of the schools within that district. Districts may receive fines or bonuses or may lose government over their schools as a result of state take-over when too many schools within the district function below some performance target. Some accountability systems also focus on a combination of levels, for example holding both individual teachers and schools accountable for student achievement. Some countries and states also instate different types of accountability systems for different types of schools. A number of states in the U.S. have for example implemented voluntary accreditation systems for regular schools and test-based accountability, including targets and stakes for title I schools.

Standards

A second component in describing accountability systems are the standards used to evaluate education and the cut points and performance targets used to assess whether schools perform adequate or not. According to Hanushek and Raymond (2001), standards involve a selection of a subset of all possible elements in a domain to both represent the whole and to be used to extrapolate more generalized performance. Standards present the details of what is expected of schools; they create boundaries or domains for attention with respect to educational quality. Usually a distinction is made in process and output/outcome standards. Process standards include activities of schools that transform input into educational outputs. These activities may for example involve classroom teaching, instructional leadership, opportunity to learn, school climate, staff development and collegial interaction among

teachers (Webster et al, 1993; Eurydice, 2004). When educational processes are evaluated, grading scales are often used, which may be applied to each standards or on a more general basis (Eurydice, 2004). Schools should for example be assessed adequately by school inspectors on standards concerning the teaching and learning in the school to receive an overall positive assessment. The study of Eurydice (2004) shows that most European accountability systems use standards on educational processes.

Output or outcome standards on the other hand include the results achieved by schools (Eurydice, 2004). They relate to the skills or knowledge that students acquire while at school. Student academic performance, teacher and student attendance rates, dropout and completion rates, performance of students at the next level of schooling, percentage of students completing advanced courses and college attendance may also be part of output standards (Webster et al, 1993). Most states in the U.S. use some form of output-based or test-based forms of accountability in which the results of achievement tests are the main source of information in evaluating education. Schools are expected to meet an absolute target (achieve a performance threshold), meet an annual growth target that is based on each school's past performance and often reflects its distance from state goals, and/ or reduce the number or percentage of students scoring in the lowest performance levels (Goertz and Duffy, 2001).

Measurement methods

Measurement methods to evaluate students, teachers, schools or districts are related to the standards in accountability systems. Educational processes are usually measured by school inspectors using protocols and procedures when visiting schools. During these visits they often interview teachers and school leaders on educational processes in the school. Sometimes students and their parents are also interviewed. School inspectors observe teachers while working with students and teaching and analyze document on for example school policy, student records and teaching materials. In some cases, the results of internal evaluations and/or peer reviews by schools are used by school inspectors to assess school. In these cases, schools are often mandated to administer internal evaluations. Lately, in some countries, school inspections are partly replaced by desk research. School inspectors use documents, facts and figures on how the school is doing (for example number of sick teachers or early plans of schools) to assess schools. Only when these documents point to weaknesses in educational quality of schools will school inspectors visit schools to complete the assessment. These so called risk-based inspections are primarily instated to increase the efficiency of school inspections and decrease administrative load of school inspections for schools.

Desk research has traditionally been the primary form of measurement in output-based accountability. States display test performance and other factors (for example absenteeism or graduation rates) as reported by schools, sometimes providing aggregations of data and judgments of performance. The state profiles of the Consortium of Policy Research in Education (2001) show that most states in the U.S., having output-based accountability systems, gather information on student absenteeism or graduation rates next to using national standardized achievement tests to measure student achievement on mathematics, reading and writing, and sometimes also science and social studies.

Stakes

The consequences or stakes attached to varying levels of performance or output may differ for distinct accountability systems. Districts, schools, teachers or students may receive bonuses or rewards for good performance or may be sanctioned when failing to meet performance targets. Some accountability systems also include interventions to help schools to improve. These interventions may consist of (state) intervention teams or experts supporting the school during a certain period of time. The most obvious stakes are the possible sanctions for schools that fail to meet the targets set in the accountability system. However, in some countries rewards and sanctions are also given to individual students and teachers. High-stakes consequences for students include for example national regulations and decisions on minimum achievement levels to be promoted to the next grade or to graduate, instead of just using test scores to determine which students need instructional help. Teachers achieving high test results sometimes receive bonuses whereas teachers with low test results are replaced by other teachers.

These sanctions and rewards may differ in type and weight depending on the extent of underperformance in schools. In the comparative study by Eurydice (2004) three types of consequences have been identified that differ in weight. The first type of consequence concerns the

recommendations or even instructions from the evaluators or education authorities. These recommendations or instructions should provide a basis for quality improvement at the school concerned. The second type of consequence is of a higher stake as it includes obligations for schools to produce a formal plan for improvement which sets out and structures the aims to be pursued. The last consequence is the most severe as it consists of disciplinary action (such as fines or withdrawal of the right to award certificates) which may be directly administered by the evaluators themselves or initiated by the education authorities. These actions may also include rewards as a result of high performance. Hamilton and Stecher (2001) mention for example money to use for school improvement or direct cash bonuses for staff. At first, external evaluators usually try to rectify low performance of schools with minor sanctions such as obliged improvement trajectories. Only when schools do not improve after some time will more severe sanctions be issued against schools.

The following table summarizes the components of accountability systems described in this paragraph. The different stakeholders in the accountability system are depicted at the upper side of the table as they are related to the other components in the accountability system.

Table 1. Components of accountability systems

		Stakeholders			
		Student level	Class level	School level	System level
Standards	Output				
	Educational processes				
Measurement methods	Tests				
	Inspection visits				
	Results of internal evaluations				
	Desk research				
Targets	Targets on educational processes				
	Targets on output (e.g. cut-score)				
Stakes	Rewards				
	Sanctions				
	Interventions				

Methods

This study is aimed at measuring the effects of different components and types of accountability systems on student achievement. An international comparison of different types of accountability systems will be used to study effectiveness of accountability systems. The methods used by Carnoy and Loeb (2002) and Hanushek and Raymond (2005) are considered to be the most methodologically sound and will be followed wherever possible with available data.

The results of studies by Eurydice (2004), the Standing conference on Inspectorate of Education (SICI) (1999) and the Consortium on Policy Research in Education (CPRE) (2001) on accountability systems in Europe and the United States, and results of studies on student achievement in mathematics (TIMSS and NAEP) are used to compare effects of different components and types of accountability systems. The sample consists of 50 states in the U.S.A. and 30 countries in Europe.

The three reports on accountability systems by Euridice, SICI and CPRE include a description of accountability systems in Europe and the states in the U.S. in 1999/2000. The Eurydice report describes approaches to external and internal evaluation of schools in European countries. The bluebook by SICI includes a description of school inspectorates in Europe, while CPRE provides state profiles describing the state assessment system, performance standards, reporting of performance data, the state accountability system and how low performing schools and districts are identified. The information in the reports was used to compose a database including the components as described in table 1. Following Carnoy and Loeb (2002), scores were given to every component in table 1 on a scale from 0-5, except for the scales on standards and targets which include the total number of standards and targets. A higher score on any scale represents a stronger (component of the)

accountability system. Strong accountability systems are, contrary to Carnoy and Loeb (2002) who only include stakes in their index of strength of accountability systems, expected to include all components of accountability systems.

The reports have been validated by the participating countries. As a result, the information provided in the reports can be considered valid and reliable. A drawback of using existing reports is however that information on some of the components may be lacking as the reports do not match the components in this study entirely. By using a missing value analysis, we tried to overcome this problem. Missing values are imputed using paired estimation. This method is considered to be the best option as missing value analysis showed data to be missing completely at random (Little's MCAR test generates a significance value of 1.000) and the overall sample size is small. All the analyses will however be run on both the original and imputed datasets to observe whether the imputation makes a difference for the results found. The results of the analyses of the original datasets are included in the appendix of this paper.

Effects of components and types of accountability systems on student achievement are measured by following one cohort of students from grade 4 to 8. Since both NAEP and TIMSS test students in mathematics in both grades 4 and 8 every four years, we can determine the gains of the cohort tested in grade 4 in 2003 and again in grade 8 in 2007. As 1999/2000 is the reference year for analyzing accountability systems, 2003 is the first year that can be used to analyze effects on test scores. Linking the results of both tests is possible as they are relatively similar in tested subjects, item format, use of grade-based samples and years of administration (Johnson et al, 2003; NCES, 2004). In some years students responded to both NAEP and TIMSS, enabling a direct comparison of performance on the two instruments. Statistical moderation (where means and standard deviations of reported distributions are matched) was used to link both tests. This method is considered to be the best option in linking NAEP and TIMSS as other methods (such as IRT) are not possible because of lacking information on item level and as other methods (such as projection-based linking where regression analysis is used to derive predictions from NAEP data about characteristics of the distribution of TIMSS) did not yield reliable results. Johnson et al (2003) used projection-based linking to estimate performance of students on TIMSS in 1999, based on their performance on NAEP in 2000. This linkage significantly under predicted actual TIMSS scores, suggesting that projection-based linking could not be used to link TIMSS and NAEP. According to these authors, statistical moderation did a good job of predicting the TIMSS scores since this linkage was based on the national data for NAEP and TIMSS and problems due to the differential functioning of TIMSS in the linking sample are not present in the moderation linkage.

As the actual students tested four years apart are not the same students, each cohort is considered to be a probability sample of their respective cohorts. The observed gain is a more or less unbiased estimate of the population gain over the period of the study. This method is found more appealing than the repeated cross-sectional design that is generally used in this type of research. It avoids distortions of characteristics of successive cohorts that are unrelated to school accountability effects but associated with performance of students on tests. This cross-sectional design will however also be used, including test scores of TIMSS/NAEP in grade 4 and 8 separately to acquire additional results.

Using regression analysis, the hypothesis is tested that accountability systems using more measurement methods, standards, targets and stakes and accountability systems that are stronger (including a large number of standards, measurement methods and stakes) make more rapid gains in student achievement and generally perform at higher levels of student achievement, holding other circumstances and policies constant. We also tested whether process-based or output-based accountability systems and accountability systems aimed at student, teacher, school or district level are more effective in terms of student achievement. Differences in circumstances and policies of each state are fixed by calculating growth in performance between fourth and eighth grades on NAEP and TIMSS and by including two explicit measures for major categories of varying country/state differences in institutional settings of the education systems (educational expenditure and pupil/teacher ratio) when calculating regression estimates.

This type of international and interstate comparison provides, according to Schütz et al (2007), the best opportunities to identify causal effects because of the variation in accountability systems between

countries and states. Accountability systems often do not vary substantially within non federal school systems, therefore an international comparison provides better opportunities to identify causal effects. Using aggregated results enables us to avoid problems of within country selectivity; it also allows capturing potential systemic effects according to Schütz, West and Wößmann (2007). The down-side of using only country-level institutional measures is however that the degrees of freedom at the country level are very limited. Because of this limitation potential interaction effects between all the components of accountability systems cannot be analyzed.

Results

This section first describes the linking of test scores of TIMSS and NAEP. Next, descriptive results are presented on the components and different types of accountability systems. The last part of this section includes results on the effectiveness of these components and types of accountability systems.

Linking TIMSS and NAEP

Mathematics test scores of students in grade 4 and 8 on TIMSS and NAEP in years 1999/2000, 2003 and 2007 were linked using statistical moderation. The following equation was used to establish the link:

$$\text{Equation 1: } \hat{T}IMSS_{level} = \hat{A} + \hat{B}(NAEP_{level}).$$

In the first equation \hat{A} is an estimate of the intercept of a straight line, and \hat{B} is an estimate of the slope defined by

$$\text{Equation 2: } \hat{B} = \frac{\hat{\mu}_{TIMSS} - \hat{B}\hat{\mu}_{NAEP}}{\hat{\sigma}_{NAEP}}$$

In equation 2, $\hat{\mu}_{NAEP}$ and $\hat{\mu}_{TIMSS}$ are the national means of the U.S. NAEP and TIMSS results for public school students, while $\hat{\sigma}_{NAEP}$ and $\hat{\sigma}_{TIMSS}$ are the standard deviations of the tests.

Following Johnson (2003), the quality of the links was estimated by comparing confidence intervals of actual and predicted means and comparing actual and predicted means of states that participated in both TIMSS and NAEP. The linkage is credible when the predicted TIMSS mean falls within the boundaries of the confidence interval of the actual mean, and when the confidence intervals of actual and predicted means overlap.

Tables 2-6 below show the actual and predicted TIMSS mean and confidence intervals for the states participating in both TIMSS and NAEP for the years and grades used in this study to measure (gains in) student achievement. The results show that there is a large agreement between the actual TIMSS results and the results predicted from NAEP. Not only do most of the confidence intervals for the predicted TIMSS means contain the actual TIMSS means and vice versa, but the intervals themselves overlap in most cases. Predicted TIMSS results from this study based on NAEP samples are about as reliable as actual TIMSS results. The linkage of TIMSS and NAEP is therefore considered to be credible.

States in the U.S. were also rank ordered on both NAEP scores and predicted TIMSS scores as a second method to observe the quality of the linking. Large differences in rank ordering point to problems in the prediction of TIMSS scores. Rank orders for NAEP and predicted TIMSS scores however showed no differences in rank orders, except for predicted TIMSS scores in 2000. Two states (Vermont and Ohio) switched places in this year. As the overall rank ordering is however almost the same, the linkage between NAEP and TIMSS scores is found to be valid.

Table 2. Linking TIMSS and NAEP mathematics grade 8 in 1999/2000

State	Actual TIMSS results mathematics grade 8 in 1999		Predicted TIMSS results mathematics grade 8 in 1999	
	Mean (SE)	95% confidence interval	Mean (SE)	95% confidence interval
Connecticut	512.00 (9.1)	493.8 – 530.2	522.63 (5.38)	511.60-533.39
Idaho	495 (7.4)	480.2 – 509.8	514.38 (4.36)	505.66-523.10
Illinois	509 (6.7)	495.6 – 522.4	508.19 (7.16)	493.87-522.51
Indiana	515 (7.2)	500.6 – 529.4	523.77 (5.84)	512.09-535.45
Maryland	495 (6.2)	480.6 – 507.4	502.23 (7.44)	487.35-517.11
Massachusetts	513 (5.9)	501.2 – 524.8	518.27 (6.22)	505.83-530.71
Michigan	517 (7.5)	502 – 532	514.60 (8.14)	498.32-530.88
Missouri	490 (5.3)	479.4 – 500.6	499.94 (6.24)	487.46-512.42
North Carolina	495 (7.0)	481 – 509	512.08 (5.50)	501.08-523.08
Oregon	514 (6.0)	502 – 526	521.02 (6.53)	507.96-534.08
South Carolina	502 (7.4)	487.2 – 516.8	485.50 (6.43)	472.64-498.36
Texas	516 (9.1)	497.8 – 534.2	505.67 (7.05)	491.57-519.77

Table 3. Linking TIMSS en NAEP mathematics grade 4 in 2003

State	Actual TIMSS results mathematics grade 4 in 2003		Predicted TIMSS results mathematics grade 4 in 2003	
	Mean (SE)	95% confidence interval	Mean (SE)	95% confidence interval
Indiana	533 (2.8)	527.4 – 538.6	528.67 (9.59)	509.49-547.85

Table 4. Linking TIMSS en NAEP mathematics grade 8 in 2003

State	Actual TIMSS results mathematics grade 8 in 2003		Predicted TIMSS results mathematics grade 8 in 2003	
	Mean (SE)	95% confidence interval	Mean (SE)	95% confidence interval
Indiana	508 (5.2)	497.6 – 518.4	515.27 (13.59)	488.09-542.45

Table 5. Linking TIMSS en NAEP mathematics grade 4 in 2007

State	Actual TIMSS results mathematics grade 4 in 2007		Predicted TIMSS results mathematics grade 4 in 2007	
	Mean (SE)	95% confidence interval	Mean (SE)	95% confidence interval
Massachusetts	572 (3.5)	565 – 579	563.52 (12)	539.52-587.52
Minnesota	554 (5.9)	542.2 – 565.8	549.50 (15.08)	519.40-579.60

Table 6. Linking TIMSS en NAEP mathematics grade 8 in 2007

State	Actual TIMSS results mathematics grade 8 in 2007		Predicted TIMSS results mathematics grade 8 in 2007	
	Mean (SE)	95% confidence interval	Mean (SE)	95% confidence interval
Massachusetts	547 (4.6)	537.8 – 556.2	545.75 (13.99)	517.77-573.73
Minnesota	532 (4.4)	523.2 – 540.8	532.96 (11.33)	510.30-555.62

Components and types of accountability systems

First of all, a table with descriptive information on the components and types of accountability systems is presented to get a first insight into the range of differences in accountability systems in Europe and the U.S. The number of components (standards, measurement methods, targets and stakes) include all the levels at which the components can be found (student, teacher, school and district level); counting the number of standards multiple times when they are used to evaluate multiple educational levels.

As can be deduced from the results in table 7, most accountability systems in Europe and the U.S. include a large number of standards and targets; the number of measurements methods and stakes is fewer. Accountability systems in Europe and the states in the U.S. vary the most in the number of standards to evaluate schools. The number of measurement methods and stakes is mostly the same in most accountability systems. Most accountability systems are aimed at evaluating schools, while students and districts are in some cases also evaluated. Accountability systems are in only a few cases designed to evaluate teachers. Most accountability systems are aimed at evaluating output of schools, while only a few systems evaluate educational processes in schools. In general, accountability systems are relatively not very strong in number of standards, targets and stakes used to evaluate students, teachers, schools and districts.

Table 7. Descriptive information of components of accountability systems

	Minimum	Maximum	Mean	Std. Deviation
Standards	.00	164.00	37.25	34.77
Measurement methods	.17	18.00	5.75	3.384
Targets	.00	54.00	7.24	10.34
Stakes	.00	19.50	6.48	4.49
Student level	.00	59.00	14.68	15.72
Teacher level	.00	8.16	.54	1.52
School level	.00	62.00	25.74	15.84
District level	.00	67.00	11.99	16.32
Output based	.00	175.00	40.41	38.14
Process based	.00	18.00	2.64	3.71
Index 'strength of accountability systems'	.00	14.92	3.69	3.28

Items related to the components and types of accountability systems were included in a reliability analysis to calculate the consistency of the scales that describe the different components and types of accountability systems. The reliability of the final scales is presented in table 8, as well as the number of items that were used. Except the scale about 'targets', all of the scales show reliable alphas. The scale on 'targets' was excluded from further analyses. The scales 'accountability systems on student level' and 'process-based accountability' have relatively low alphas. Given the few number of items in these scales, the alphas are considered to be acceptable and the scales are included in further analyses. Table A1 in the appendix presents the reliability analysis for the original data including missing values. The alphas in this table are somewhat higher, but do not imply other conclusions.

Table 8. Main results of reliability analyses on different components in accountability systems

	Cronbach's Alpha, imputed database	Number of items
<i>Standards</i> Number of standards on output and educational processes (class-level and school level) to evaluate students, teachers, schools and local communities or districts	.683	7
<i>Measurement methods</i> Types of measurement methods used to evaluate students, teachers, schools and local communities or districts	.763	11
<i>Targets</i> Output and process-related targets on student, teacher, school and local community or district level	-0.80	5
<i>Stakes</i> Rewards, sanctions and interventions on student, teacher, school and local community or district level	.696	12
<i>Accountability system on student level</i> Standards, measurement methods, targets and stakes aimed at students	.456	5
<i>Accountability system on teacher level</i> Standards, measurement methods, targets and stakes aimed at teachers	.660	8
<i>Accountability system on school level</i> Standards, measurement methods, targets and stakes aimed at schools	.805	8
<i>Accountability system on local community/district level</i> Standards, measurement methods, targets and stakes aimed at local communities/districts	.675	5
<i>Output-based accountability systems</i> Standards, measurement methods (tests), targets and stakes aimed at output of students, teachers, schools, local communities or districts	.771	5
<i>Process-based accountability systems</i> Standards, measurement methods (tests), targets and stakes aimed at educational processes on teacher, school, local community or district level	.586	3
<i>Index strength of accountability systems</i> Number of standards, measurement methods, targets and stakes to evaluate students, teachers, schools and local communities and districts	.726	12

Effectiveness of components and types of accountability systems

This study concentrates on the effect of different components and types of accountability systems on (gains in and levels of) student achievement in mathematics. First, correlations were measured to acquire a basic overview of relations between accountability and student achievement. The components (standards, measurement methods and stakes) and levels (student, teacher, school and district) of accountability systems, output-based versus process-based accountability and the index 'strength of accountability systems' were related to gains in math achievement and average math achievement levels of countries and states in grade 4 and 8 in 2003 and 2007. The results in table 9 show only significant correlations for measurement methods and student level and school level accountability systems with average mathematics achievement in grade 4 in 2007. These results indicate that accountability systems situated at student level and school level, using a larger number of measurement methods have a positive impact on student achievement in mathematics in grade 4 in 2007. However, a Z-score was calculated to estimate the probability that these correlations are biased by having performed many tests on the same data. The Z-score is 0.9733, pointing to a significance level of 0.1652. There is a 16.52% chance that these correlations are based on chance instead of pointing to actual relationships in the data. As a significance level of 0.05 or 5% is generally accepted, these results can only be used exploratory as indications of potential relevant components of accountability systems.

Table 9. Correlations between components and types of accountability systems and output and gains

	Gains: 2003 grade 4 math achievement to 2007 grade 8 math achievement	Average math achievement grade 4 2003	Average math achievement grade 8 2003	Average math achievement grade 4 2007	Average math achievement grade 8 2007
Standards	-.003	-.040	-.104	.038	-.040
Measurement methods	.103	.050	-.056	.267(*)	.103
Stakes	.013	.057	-.139	.215	.062
Student level accountability	-.026	.091	-.056	.231(*)	.075
Teacher level accountability	-.148	-.109	-.252(*)	-.102	-.185
School level accountability	.125	.096	-.053	.277(*)	.160
District level accountability	.049	.057	.031	.130	.082
Output based accountability	.024	.001	-.131	.160	.014
Process based accountability	-.118	-.005	.077	-.176	-.068
Index 'strength of accountability systems'	.025	.017	-.120	.183	.030

* Correlation is significant at the 0.05 level (2-tailed)

** Correlation is significant at the 0.01 level (2-tailed)

Next, linear regression analysis was used to predict gains in student achievement and levels of student achievement based on components and types of accountability systems. Gains include differences in math achievement of students tested in grade 4 in 2003 and again in grade 8 in 2007. Educational expenditure (percentage of Gross Domestic Product on public (pre)primary, secondary and post secondary education) and pupil-teacher ratio per country and state, and student achievement in mathematics in 2000 were included in the regression framework to correct for other types of differences in policies and circumstances in countries and states and to correct for prior achievement. These covariables were the only relevant data available on both European countries and states in the U.S.

The regression framework first includes the basic model of covariables. In model two to five, different components and types of accountability systems are added, starting with the components (standards, measurement methods and stakes); next, the levels of accountability systems (student, teacher, school, district) are added to the basic model. Model four adds process-based versus output-based accountability systems to the basic model. In the last model the index 'strength of accountability systems' is included.

The results of the regression analyses indicate that only accountability systems on school level (model 3) significantly affect gains in student achievement. States and countries that have accountability systems aimed at schools achieve 4.81 points more gains in mathematics achievement. To put these gains in perspective, on average the variable 'gains' has a standard deviation of 10.08. Other components of accountability systems, such as standards, measurement methods and stakes do not seem to contribute to gains in student achievement. Also, strong accountability systems and

differences in output or process based accountability systems do not explain gains in student achievement. Similar results were found in the original data (see table A3 in the appendix).

As the current accountability policies are mostly aimed at performance at the bottom end of the achievement distribution, Hanushek and Raymond (2005) suggest that an alternative would be to analyze performance at other points in the distribution. The regression analyses were therefore also computed for gains in student achievement at the 25th, 50th and 75th percentile of the distribution. The results are however not significant and are not included in this paper.

Table 10. Strength of accountability systems related to gains in educational performance

Independent variable	Model 1 (basic)	Model 2 (components)	Model 3 (levels)	Model 4 (output vs. process-based)	Model 5 (index)
Average student achievement in mathematics grade 8 1999/2000	0.077 (1.486)	.086 (1.582)	.099 (1.837)	.089 (1.621)	.088 (1.619)
Percentage of GDP on public (pre)primary, secondary and post secondary education	3.338 (1.961)	2.941 (1.678)	3.317 (1.940)	3.036 (1.732)	3.108 (1.787)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.451 (-.972)	-.591 (-1.232)	-.542 (-1.149)	-.440 (-.896)	-.514 (-1.084)
Standards		.009 (.128)			
Measurement methods		4.836 (1.281)			
Stakes		-1.591 (-.438)			
Student level accountability			-1.475 (-1.218)		
Teacher level accountability			-3.913 (-.781)		
School level accountability			4.807* (2.060)		
District level accountability			-.564 (-.477)		
Output based accountability				.056 (.289)	
Proces based accountability				-.962 (-.576)	
Index 'strength of accountability systems'					.252 (.691)
Constant	-54.795 (-1.792)	-56.908 (-1.797)	-67.509* (-2.154)	-59.638 (-1.888)	-59.412 (-1.892)
R square	.098	.120	.163	.107	.104
Sample size ²	81	81	81	81	81

Note: dependent variable: gains: 2003 grade 4 math to 2007 grade 8 math (TIMSS/NAEP); *t*-statistic in parentheses, * $p < .05$, ** $p < .01$

The same regression analyses, using multivariate regression, were computed using output levels of average student achievement. The impact of educational expenditure, pupil-teacher ratio and prior achievement (basic model), components of accountability systems, levels of accountability systems, output versus process based accountability and strength of accountability on output levels of student achievement in mathematics in grade 4 and 8 in 2003 and 2007 were measured. The results in table 11

² The sample size is 81 (instead of 80) as a result of including the U.S. average in the analyses

point to an impact of particularly the index 'strength of accountability systems', school level accountability systems and output based accountability systems on (most) output levels of student achievement in mathematics. Particularly school level accountability has a high impact, respectively 11.23 and 11.58 points in student achievement in grade 4 and grade 8 in 2007. Measurement methods in accountability systems have a relatively high impact (13.55) on grade 4 student achievement in math in 2007, whereas accountability systems aimed at teachers have a surprisingly negative impact on student achievement in mathematics in grade 8 in 2003. Other components and types of accountability systems do not affect student achievement levels.

The results of the analyses on the original data lead to somewhat other conclusions. Table A4 in the appendix shows only a positive impact of school level accountability systems and stakes in accountability systems on student achievement in respectively mathematics achievement grade 8 in 2007 and grade 4 in 2003. Measurement methods in accountability systems even have a negative impact on math achievement in grade 4 in 2003.

Table 11. Strength of accountability systems related to output in educational performance

	Dependent variables			
	Average math achievement grade 4, 2003	Average math achievement grade 8, 2003	Average math achievement grade 4, 2007	Average math achievement grade 8, 2007
Independent variables:				
Model 1:				
Average student achievement in mathematics grade 8 1999/2000	.450** (5.335)	.636** (10.074)	.371** (3.803)	.527** (6.458)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-3.019 (-1.095)	-1.741 (-.843)	3.197 (1.003)	.318 (.119)
Pupil/teacher ratio for pre-primary, primary and secondary education	.038 (.051)	-.291 (-.516)	.561 (.646)	-.413 (-.567)
Constant	297.650** (6.007)	192.831** (5.193)	313.771** (5.478)	242.854** (5.062)
R square	.295	.592	.162	.374
Sample size	81	81	81	81
Model 2:				
Average student achievement in mathematics grade 8 1999/2000	.510** (5.986)	.665** (10.232)	.462** (4.928)	.596** (7.444)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-4.489 (-1.640)	-2.655 (-1.271)	.723 (.240)	-1.548 (-.601)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.293 (-.391)	-.552 (-.966)	-.076 (-.092)	-.885 (-1.255)
Standards	-.006 (-.060)	-.003 (-.037)	0.020 (.175)	.002 (.024)
Measurement methods	5.107 (.866)	7.278 (1.617)	13.548* (2.087)	9.943 (1.793)

	Dependent variables			
	Average math achievement grade 4, 2003	Average math achievement grade 8, 2003	Average math achievement grade 4, 2007	Average math achievement grade 8, 2007
Stakes	9.139 (1.612)	1.013 (.234)	9.388 (1.505)	7.548 (1.416)
Constant	269.699** (5.451)	182.037** (4.823)	274.245** (5.036)	212.790** (4.574)
R square	.360	.616	.310	.464
Sample size	81	81	81	81
Model 3:				
Average student achievement in mathematics grade 8 1999/2000	.513** (6.053)	.665** (10.809)	.473** (5.078)	.612** (7.984)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-3.591 (-1.462)	-2.343 (-1.195)	1.869 (.629)	-.634 (-.260)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.406 (-.546)	-.476 (-.880)	-.145 (-.177)	-.948 (-1.409)
Student level accountability	1.374 (.718)	.003 (.002)	1.797 (.854)	-.101 (-.059)
Teacher level accountability	-1.445 (-.183)	-14.460* (-2.517)	2.215 (.255)	-5.359 (-.749)
School level accountability	6.772 (1.837)	4.627 (1.729)	11.231** (2.771)	11.579** (3.475)
District level accountability	.251 (.134)	-.017 (-.012)	.217 (.106)	-.313 (-.185)
Constant	267.122** (5.393)	179.574** (4.994)	263.163** (4.834)	199.612** (4.460)
R square	.377	.662	.331	.518
Sample size	81	81	81	81
Model 4:				
Average student achievement in mathematics grade 8 1999/2000	.510** (5.984)	.679** (10.464)	.459** (4.719)	.599** (7.294)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-4.247 (-1.541)	-2.549 (-1.232)	1.207 (.389)	-1.211 (-.463)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.252 (-.327)	-.572 (-.988)	.337 (.389)	-.692 (-.944)
Output based accountability	.586 (1.933)	.482* (2.122)	.686* (2.013)	.641* (2.230)
Process based accountability	-.357 (-.136)	.727 (.369)	-3.193 (-1.080)	-1.320 (-.529)
Constant	272.526** (5.484)	174.832** (4.688)	277.015** (4.955)	212.888** (4.512)

	Dependent variables			
	Average math achievement grade 4, 2003	Average math achievement grade 8, 2003	Average math achievement grade 4, 2007	Average math achievement grade 8, 2007
R square	.342	.618	.259	.438
Sample size	81	81	81	81
Model 5:				
Average student achievement in mathematics grade 8 1999/2000	.513** (6.018)	.682** (10.635)	.462** (4.796)	.602** (7.437)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-4.330 (-1.595)	-2.682 (-1.314)	1.313 (.428)	-1.222 (-.474)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.325 (-.438)	-.551 (-.989)	.040 (.048)	-.839 (-1.194)
Index	1.438* (2.528)	1.033* (2.413)	2.068** (3.218)	1.690** (3.131)
Constant	271.284** (5.534)	173.904** (4.718)	275.872** (4.984)	211.872** (4.555)
R square	.350	.621	.262	.445
Sample size	81	81	81	81

Note: *t*-statistic in parentheses; * $p < .05$, ** $p < .01$

The regression analyses were also computed for output levels in student achievement in mathematics at the 25th, 50th and 75th percentile of the distribution to see if accountability systems have an effect on specific points in the distribution of student achievement. The results show that *teacher level accountability* has a significant effect on student achievement at the 25th percentile of the distribution in math in grade 4 in 2003 and 2007 and grade 8 in 2007 and with student achievement at the 75th percentile of the distribution in grade 8 in 2007. *School level accountability* has a significant effect on student achievement in math in grade 4 in 2007 at the 25th percentile, with math student achievement in grade 4 and 8 in 2007 at the 50th percentile and with student achievement in grade 4 and 8 in 2007 at the 75th percentile of the distribution. *Stakes in accountability systems* significantly affect math student achievement in grade 4 in 2003 and grade 4 and 8 in 2007 at the 75th percentile level of the distribution of student achievement. The tables are not included here for purposes of readability.

Conclusion and discussion

Educational accountability has been part of education some time now. A number of European countries have a long tradition of accountability systems; some school inspections date back some 200 years ago. In the U.S., attention on school accountability has increased considerably since the enactment of the No Child Left Behind Act in 2001. States were mandated to introduce accountability systems including annual testing of all students in grade 3 through 8 by 2006 and to disaggregate data on student performance for all schools. States were also mandated to develop performance targets for schools along with a variety of sanctions if schools fail to meet these targets. As a result of these measures a lot of research in the U.S. has focused on measuring effects and side effects of high stakes test-based accountability; whereas research efforts in Europe have traditionally focused on effectiveness of school inspections and other topics related to school inspections (such as reliability of performance measurement). Since these two types of accountability systems are becoming more and more intertwined, research into accountability systems should also use an integrative approach. The study presented in this paper is an example of such an approach. General components of accountability

systems, such as measurement methods, standards, targets and stakes, were deduced to describe different types of accountability systems ranging from process-based to output-based accountability, differing in educational levels accountability systems aim at (student, teacher, school, district) and differing in strength of accountability systems. This approach enables a more detailed study of the specific types of accountability systems and elements within these systems. The study presented in this paper is aimed at finding out the impact of these different elements and types of accountability systems on student achievement.

The results of this study imply that the presence of accountability systems alone matters, particularly when accountability systems evaluate schools and include incentives to stimulate high performance in schools. Accountability systems aiming at schools have a positive impact on both gains in student achievement in math from grade 4 to grade 8 and output levels in student achievement. The effect of school level accountability is small but, considering the fact that student achievement in the U.S. and Europe is relatively of the same level, this effect is still relevant.

Effective (school level) accountability systems include a relatively large number of standards, measurement methods and stakes. The structure and specific type of accountability systems seem irrelevant as no convincing effects were found for particular components of accountability systems or for output-based versus process-based accountability systems. Accountability systems aiming at students, teachers or districts (next to, or instead of schools) also have no impact on student achievement. This conclusion may however also be a reflection of limited variance found in the levels of accountability systems studied here. Most accountability systems in this study are situated (only at) the school level.

As most accountability systems in Europe and the United States are aimed at educational quality provided by schools, the findings presented here underscore the importance of most accountability systems. This however raises the question whether a person or an organization should be held solely accountable for matters involving a shared responsibility or for matters that are affected by factors outside of the person's or organization's control. As stated by Leithwood and Earl (2000), the success of students in schools, for example, is a function of many factors. Students have to actively participate in the learning and teaching activities they are offered, teachers have to offer high quality instruction and guidance in their classes, school principals have to organize educational processes that support the teaching and learning in their schools and districts or local communities have to facilitate the schools and school boards in their region. Also, other factors such as family educational culture of students accounts for a large part of the variation in student achievement. According to Leithwood and Earl (2000) it would be legitimate to hold the teacher or principal accountable for making the most productive uses of the resources available to them in an effort to move toward the goal, instead of holding them solely accountable for actually achieving the goal. Approaches to accountability should be distinguished by the nature of the obligation a person or group is assumed to have and the extent to which that obligation is legitimate. Administrators who implement accountability systems should take these responsibilities and obligations into consideration in designing standards, targets, measurement methods and stakes.

The results of this study still leave room for a lot of questions such as why accountability systems actually work and if these systems positively affect student achievement on both the short and long run. According to Schildkamp and Ehren (in press) inspection data may ultimately result in greater awareness of data to improve educational processes and output of the school because of the constant focus of school inspectors on use of data to improve educational processes and output of schools. Ehren and Visscher (2008) also found that performance feedback of school inspectors and assessments of educational processes and output of schools lead to school improvement. The capacity of schools to change and external pressure or support from their environment are considered to be important conditions in these improvement processes. These elements may also apply to test-based accountability systems.

Further research should try to reveal the behavioral mechanisms explaining effects of accountability systems. Additional research should for example investigate the period of time accountability systems have an effect on educational quality and the activities schools employ that lead to effects. As test scores are known to increase after new assessments and accountability provisions are put into place, the start dates of accountability systems should be taken into consideration to see if effects of

accountability systems wear out after a few years. If this is the case, special attention should also be paid to potential side effects of accountability systems such as test inflation. In the case of score inflation, test scores increase while students' knowledge in the tested domain does not increase. Increases in test scores are actually the result of undesirable methods of schools such as undesirable test preparation, educational triage or fraud. A method to overcome this problem is using tests that are not part of accountability systems, such as NAEP and TIMSS in this study, to measure effects of accountability systems. Schools have no stake in achieving high test scores on these tests and are expected to refrain from acting towards side effects.

Further research should also try to include a wider variability of accountability systems, for example accountability systems situated at student or teacher levels, to study potential effects of these types of systems. Variations within a state or country (in for example the specific number of rewards and sanctions given to schools) may also provide valuable insights into effectiveness of (components of) accountability systems. Countries or states having different accountability systems for different types of schools may serve as useful research objects in such a design. Studies within one school system also enable inclusion of school factors, such as socio-economic backgrounds of the pupil population or prior performance of schools. Prior research (Jacob and Levitt, 2003; Stecher, 2001) has indicated that these school factors may explain both effects and side effects of accountability systems. In the end, results of these studies should contribute to the design of more effective accountability systems, stimulating and enabling high achievement of students, teachers, schools and administrators.

References

- Anderson, J.A. (2005). *Accountability in Education*. Paris/Brussels: Unesco.
- Berry, B., Turchi, L., Johnson, D., Hare, D., Duncan Owens, D. (2003). *The Impact of High-Stakes Accountability on Teachers' Professional Development: Evidence from the South*. A Final report to the Spencer Foundation.
- Booher-Jennings, J. (2005). Below the Bubble: 'Educational Triage' and the Texas Accountability System. *American Educational Research Journal*, 42(2), 231-268.
- Carnoy, M. and Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Chapman, C. (2001). Changing classrooms through inspection. *School leadership and management*, 21(1), 59-73.
- Consortium for Policy Research in Education (CPRE) (2001). *Accountability and Assessment Profiles*. http://www.cpre.org/index.php?option=com_content&task=view&id=169&Itemid=124
- Council of Chief State School Officers (CCSSO) (2004). *A Framework for Examining Validity in State Accountability systems*. Washington: CCSSO.
- Ehren, M.C.M. (2006). *Toezicht en schoolverbetering*. Delft: Uitgeverij Eburon
- Ehren, M.C.M., Visscher, A.J. (2008). The Relationship between School Inspections, School Characteristics and School Improvement. *British Journal of Educational Studies*, 56(2), 205-227.
- Eurydice (2004). *Evaluation of Schools providing Compulsory Education in Europe*. <http://www.eurydice.org/portal/page/portal/Eurydice>
- Goertz, M.E. and Duffy, M.C. (2001). *Assessment and Accountability Across the 50 States*. Policy brief RB-33. http://www.cpre.org/images/stories/cpre_pdfs/rb33.pdf
- Gribben, M.A., Campbell, H.L. and Mathew, J. (2008). Are Advanced Students Advancing? Examining Achievement Trends Beyond Proficiency. *Paper presented at AERA 2008*.
- Hamilton, L.S. and Stecher, B.M. (2001). Improving test-based accountability. In Hamilton, L.S., Stecher, B.M., Klein, S.P. (Eds.). *Making sense of Test-based Accountability in Education*. Santa Monica: Rand cooperation. http://www.rand.org/pubs/monograph_reports/MR1554/
- Hanushek, E.A. and Raymond, M.E. (2001). The Confusing World of Educational Accountability. *National Tax Journal*, 54(2), 365-384.
- 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.
- Hanushek, E.A. and Raymond, M.E. (2005). Does School Accountability Lead to Improved Student Performance? *Journal of Policy Analysis and Management*, 24(2), 297-327.
- Jacob, B.A. and Levitt, S.D. (2003). Rotten Apples: an investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics (august)*, 843-877.

Johnson, E., Cohen, W.H., Jiang, T. Zhang, Y. (2003). 2000 NAEP-1999 TIMSS Linking Report (NCES 2005-01). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Koretz, D. (2002). Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *The Journal of Human resources: education, manpower and welfare*, 37(4), 752-777.

Koretz, D.M. (2003). Using Multiple Measures to Address Perverse Incentives and Score Inflation. *Educational Measurement*, 22(2), 18-26.

Leithwood, K. & Earl, L. (2000). Educational Accountability Effects: An International Perspective. *Peabody Journal of Education*, 75(4), 1-18.

Luginbuhl, R., Webbink, D. & De Wolf, I. (2007). *Do School Inspections Improve Primary School Performance?* CPB-paper, nr. 83. <http://www.cpb.nl/nl/pub/cpbreeksen/discussie/83/disc83.pdf>

National Center for Education Statistics in the US (2004). *Comparing NAEP, TIMSS, and PISA in Mathematics and Science*. http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf

Rosenthal, L. (2004). Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review*, 23(2), 143-152.

Schildkamp, K. & Ehren, M.C.M. (in press). *An Exploratory Study into the Use of Accountability Data in the Netherlands*.

Schütz, G., West, M.R., and Wößmann, L. (2007). *School Accountability, Autonomy, Choice, and the Equity of Student Achievement: International Evidence from PISA 2003*. OECD: EDU/WKP(2007)9

Shaw, I, Newton, D.P., Aitkin, M. & Darnell, R. (2003). Do OFSTED inspections of secondary education make a difference to GCSE results? *British Educational Research Journal*, 29(1), 63-76.

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.

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

Webster, W.J., Mendro, R.L. & Almaguer, T.O. (1993). Effectiveness Indices: The Major Component of An Equitable Accountability System. *Paper presented at AERA 1993*.

Wolf, I.F. de, Janssens, F.J.G. (2005). *Effects and side effects of inspections and accountability in education; an overview of empirical studies*. <http://www1.fee.uva.nl/scholar/wp/wp53-05.pdf>

AppendixTable A1. Main results of reliability analyses on different components in accountability systems, original database.

	Cronbach's Alpha	Number of items
<i>Standards</i> Number of standards on output and educational processes (class-level and school level) to evaluate students, teachers, schools and local communities or districts	.679	7
<i>Measurement methods</i> Types of measurement methods used to evaluate students, teachers, schools and local communities or districts	.736	11
<i>Targets</i> Output and process-related targets on student, teacher, school and local community or district level	-0.80	5
<i>Stakes</i> Rewards, sanctions and interventions on student, teacher, school and local community or district level	.725	12
<i>Accountability system on student level</i> Standards, measurement methods, targets and stakes aimed at students	.453	5
<i>Accountability system on teacher level</i> Standards, measurement methods, targets and stakes aimed at teachers	.784	8
<i>Accountability system on school level</i> Standards, measurement methods, targets and stakes aimed at schools	.794	8
<i>Accountability system on local community/district level</i> Standards, measurement methods, targets and stakes aimed at local communities/districts	.758	5
<i>Output-based accountability systems</i> Standards, measurement methods (tests), targets and stakes aimed at output of students, teachers, schools, local communities or districts	.764	5
<i>Process-based accountability systems</i> Standards, measurement methods (tests), targets and stakes aimed at educational processes on teacher, school, local community or district level	.742	3
<i>Index strength of accountability systems</i> Number of standards, measurement methods, targets and stakes to evaluate students, teachers, schools and local communities and districts	.709	12

Table A2. Correlations between components and types of accountability systems and output and gains (original database)

	Gains: 2003, grade 4 math to 2007, grade 8	Average math achievement grade 4 2003	Average math achievement grade 8 2003	Average math achievement grade 4 2007	Average math achievement grade 8 2007
Standards	-.154	-.180	-.219	-.062	-.115
Measurement methods	-.120	-.180	-.226	.021	-.079
Stakes	-.190	.037	-.189	.198	.003
Student level accountability	-.158	-.027	-.080	.111	.054
Teacher level accountability	-.098	-.070	-.148	-.019	-.089
School level accountability	-.061	-.062	-.186	.148	.034
District level accountability	-.166	-.042	-.092	.100	.045
Output based accountability	-.160	-.145	-.170	.009	-.031
Process based accountability	-.031	.165	.116	.018	-.088
Index 'strength of accountability systems'	-.165	-.173	-.217	-.037	-.100

** Correlation is significant at the 0.01 level (2-tailed).

Table A3. Strength of accountability systems related to gains in educational performance (TIMSS/NAEP) (original database)

Independent variable	Model 1 (basic)	Model 2 (components)	Model 3 (levels)	Model 4 (output vs. process-based)	Model 5 (index)
Average student achievement in mathematics grade 8 1999/2000	.083 (1.427)	.067 (.987)	.130 (2.009)	.105 (1.593)	.090 (1.389)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-.093 (-.047)	1.545 (.679)	1.559 (.768)	.242 (.102)	-.429 (-.202)
Pupil/teacher ratio for pre-primary, primary and secondary education	.020 (.042)	.258 (.485)	.521 (1.059)	-.238 (-.417)	.110 (.213)
Standards		-.098 (-.353)			
Measurement methods		4.161 (.960)			
Stakes		-2.310 (-.668)			
Student level accountability			-1.547 (-1.651)		
Teacher level accountability			5.327 (.489)		
School level accountability			7.288* (2.704)		
District level accountability			-3.137 (-1.419)		
Output based accountability				.288 (1.324)	
Process based accountability				.593 (.184)	
Index 'strength of accountability systems'					.102 (.248)
Constant	-51.380 (-1.589)	-53.427 (-1.501)	-95.537* (-2.660)	-62.226 (-1.798)	-55.182 (-1.596)
R square	.051	.092	.258	.101	.053
Sample size	43	39	39	38	42

Note: dependent variable: gains: 2003 grade 4 math to 2007 grade 8 math (TIMSS/NAEP); *t*-statistic in parentheses, * $p < .05$, ** $p < .01$

Table A4. Strength of accountability systems related to output in educational performance (TIMSS/NAEP) (original database)

	Dependent variables			
	Average math achievement grade 4, 2003	Average math achievement grade 8, 2003	Average math achievement grade 4, 2007	Average math achievement grade 8, 2007
Independent variables:				
Model 1:				
Average student achievement in mathematics grade 8 1999/2000	.636** (9.608)	.810** (17.142)	.571** (6.587)	.719** (9.705)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-.936 (-.412)	-1.045 (-.645)	2.340 (.788)	-1.029 (-.405)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.625 (-1.120)	-.635 (-1.598)	-.472 (-.647)	-.604 (-.968)
Constant	210.999** (5.720)	111.355** (4.233)	239.172** (4.959)	159.619** (3.868)
R square	.725	.892	.563	.727
Sample size	43	43	43	43
Model 2:				
Average student achievement in mathematics grade 8 1999/2000	.666** (10.331)	.830** (14.809)	.565** (6.301)	.733** (9.308)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-2.417 (-1.117)	-2.080 (-1.105)	3.596 (1.195)	-.827 (-.330)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.410 (-.811)	-.403 (-.916)	-.049 (-.070)	-.153 (-.247)
Standards	.106 (.399)	.202 (.878)	-.280 (-.761)	.007 (.023)
Measurement methods	-8.736* (-2.119)	-3.994 (-1.114)	-4.041 (-.705)	-4.575 (-.909)
Stakes	9.756** (2.967)	3.746 (1.310)	7.860 (1.719)	7.446 (1.854)
Constant	194.494** (5.744)	99.278** (3.371)	227.771** (4.839)	141.068** (3.411)
R square	.830	.905	.675	.790
Sample size	39	39	39	39
Model 3:				
Average student achievement in mathematics grade 8 1999/2000	.659** (8.778)	.844** (15.186)	.567** (5.790)	.789** (10.121)
Percentage of GDP on public (pre)primary,	-.930 (-.384)	-1.058 (-.590)	4.350 (1.379)	.669 (.266)

	Dependent variables			
	Average math achievement grade 4, 2003	Average math achievement grade 8, 2003	Average math achievement grade 4, 2007	Average math achievement grade 8, 2007
secondary and post secondary education				
Pupil/teacher ratio for pre-primary, primary and secondary education	- .865 (-1.514)	-.513 (-1.213)	-.541 (-.726)	-.345 (-.580)
Student level accountability	1.122 (1.030)	.375 (.466)	.691 (.487)	-.425 (-.376)
Teacher level accountability	-1.256 (-.099)	1.241 (.133)	-7.766 (-.471)	4.071 (.310)
School level accountability	-2.280 (-.089)	1.669 (.720)	1.918 (.470)	7.009* (2.156)
District level accountability	.089 (.035)	-.988 (-.520)	-1.924 (-.575)	-3.048 (-1.143)
Constant	201.252** (4.823)	89.780** (2.907)	231.394** (4.254)	105.715* (2.440)
R square	.800	.917	.657	.818
Sample size	39	39	39	39
Model 4:				
Average student achievement in mathematics grade 8 1999/2000	.665** (9.430)	.834** (14.798)	.640** (8.202)	.770** (9.635)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-2.207 (-.867)	-1.607 (-.790)	-1.597 (-.568)	-1.965 (-.681)
Pupil/teacher ratio for pre-primary, primary and secondary education	-1.139 (-1.870)	-.737 (-1.515)	-1.851 (-2.749)*	-1.376 (-1.995)
Output based accountability	-.007 (-.028)	.131 (.705)	-.009 (-.035)	.281 (1.068)
Proces based accountability	-2.194 (-.635)	-2.241 (-.812)	3.951 (1.034)	-1.601 (-.409)
Constant	210.282** (5.686)	102.581** (3.472)	241.313** (5.899)	148.056** (3.532)
R square	.788	.894	.760	.783
Sample size	38	38	38	38
Model 5:				
Average student achievement in mathematics grade 8 1999/2000	.649** (9.442)	.831** (15.757)	.567** (6.087)	.738** (9.049)
Percentage of GDP on public (pre)primary, secondary and post secondary education	-.820 (-.363)	-1.562 (-.900)	2.855 (.931)	-1.249 (-.465)
Pupil/teacher ratio for pre-primary, primary and secondary education	-.960 (-1.751)	-.728 (-1.729)	-.753 (-1.013)	-.851 (-1.306)

	Dependent variables			
	Average math achievement grade 4, 2003	Average math achievement grade 8, 2003	Average math achievement grade 4, 2007	Average math achievement grade 8, 2007
Index	.189 (.431)	.326 (.969)	-.050 (-.084)	.291 (.559)
Constant	208.132** (5.652)	102.361** (3.621)	243.006** (4.867)	152.950** (3.498)
R square	.764	.894	.598	.740
Sample size	42	42	42	42

Note: *t*-statistic in parentheses, * $p < .05$, ** $p < .01$