

## **Effective use of self-evaluations of schools in external accountability systems in education**

*Melanie C.M. Ehren, University of Twente, the Netherlands*

### **Abstract**

This paper describes accountability systems using results of self-evaluations of schools to assess educational quality in schools. These accountability systems are expected to be most effective as quality assurance in schools is stimulated through external accountability. This assumption was studied, using a database on accountability systems in Europe and the U.S. and student achievement results as measured in TIMSS and NAEP. The results show how self-evaluations are used in external accountability. The results also show that the use of self-evaluations in external accountability in general does not seem to be effective in terms of improving student achievement and may even be ineffective. The results are explained by analyzing the way self-evaluations are used in accountability systems.

Keywords:

Accountability, Self-evaluations, External evaluations, School evaluation

### **Introduction**

Quality assurance, self evaluation of schools, external evaluations and accountability receive increasing attention in education. They are seen as important means to measure outcomes and school quality and as means to enhance the school's capacity to change internal processes in order to achieve educational goals more effectively. The decentralization of most educational systems underlines the significance of self evaluations in schools and external accountability as educational quality is no longer regulated by means of top down policy. (Scheerens, van Amelsfoort and Donoughue, 1999).

Janssens and van Amelsfoort (2008, p.16) describe quality assurance as the active focus of schools in ensuring the quality of the education they provide and, if possible, improving its quality. School self evaluation is part of quality assurance and includes a procedure which is initiated and carried out by the school in order to describe and evaluate its own functioning; it involves the systematic gathering of information using a procedure that is initiated by the school itself (Vanhoof, Van Petegem, De Maeyer, 2008). The information is used to assess the functioning of the school and the attainment of its educational goals for purposes of supporting decision-making and learning and for fostering school improvement as a whole (Janssens and Van Amelsfoort, 2008; Vanhoof and Van Petegem, 2009). External evaluations are predominantly used to provide a public account of the quality of schools and the education they provide, according to Janssens and Van Amelsfoort (2008). Accountability refers to holding public institutions and services responsible for the quality and output of their performance (Scheerens, van Amelsfoort and Donoughue, 1999, p.81).

Even though the basic orientations of self-evaluations and external accountability differ (improving schools versus holding schools to account), most governments also regard external evaluation as a vehicle for school improvement. External and internal self-evaluations are used as two interrelated areas of one strategy of school reform and school improvement. As a result, many accountability systems use results of self evaluations of schools in external accountability systems to achieve both goals. Barber (2004) and Janssens and Van Amelsfoort (2008) for example describe proportional inspections that take self-evaluations as a starting point. The frequency and the form of school inspections are based on the quality of the school and the risks of decline in quality. The Inspectorate of Education will not re-investigate aspects of the quality of education that have already been evaluated properly by the school. As a result, school inspections will require for fewer inspector days when schools annually conduct a thorough self-evaluation, using external accountability standards. Only where the inspector questions the effectiveness of the self-evaluation will a much fuller external inspection visit follow. Results of internal self-evaluations are also used in output or test-based accountability systems when schools are required to evaluate their educational quality by administering national standardized tests to their pupils to gather insight in performance of their pupils with respect to some national output target; schools are sanctioned when their overall test scores are too low or they receive (financial) rewards when they have achieved high scores.

According to Barber (2004), the major underpinning assumption of this combination of internal self-evaluations and external accountability is that the accountability and education system has reached a new level of maturity where performance improvement no longer needs to be driven from top down. Instead, school leaders and teachers are expected to have the will and the means to improve education on their own and are given more autonomy in making decisions on the functioning of their school to do so. Autonomous schools are urged to define their own educational aims, to be in charge of the education process and to evaluate their actions. With this expansion of school autonomy, schools are however also expected to take greater responsibility and to be accountable for their deeds according to Nevo (2001). Politicians, administrators, parents and the public want to know if schools are fulfilling their duties. Using self evaluations of schools in quality controls by government is chosen as a way of responding to these school accountability requirements and, at the same time leaving schools in charge of their educational aims and ways to improve. According to Nevo (2001), internal self-evaluation can also deepen the scope of external evaluation by increasing its awareness of local issues. It can improve the interpretation of external findings by making them more sensitive to local needs.

Governments however also use external accountability systems to provide some strategic direction to schools by motivating schools to address certain standards in their self evaluations. Goertz and Massell (in Grooss and Goertz, 2005) describe how schools, departments and individual teachers, as a result of external accountability, adopt a plethora of accountability-related initiatives, from voluntary tutoring sessions or test preparation activities to more comprehensive overhauls of curriculum and instruction. They describe how educators across the states have come to accept the fundamental premise of standards, and see the aim of common, strong academic goals as compatible with good educational practice (Goertz and Massell, in Grooss and Goertz, 2005). Accountability systems using internal self-evaluations are also expected to be effective as quality assurance and school improvement is stimulated through external evaluations and assessments. Experience shows that schools might be motivated to engage in self-evaluation if faced with an external evaluation requirement, even when internal self-evaluation is not suggested as an alternative to external evaluation but only as a prior condition and counterpart. Internal self evaluations are expected to provide schools with the means to improve decision-making processes and to make schools more effective. According to Nevo (2001), schools that have internal self-evaluation teams and have incorporated evaluations in their pedagogical and administrative life, have a better chance of understanding the meaning of evaluation and its significance in education.

Using self evaluations as part of external accountability may however also lead to undesirable side effects. Janssens (2005) describes how schools may be inclined to withhold on reporting about risky decisions and failures and weak points in the school in self evaluations when a third party reads these evaluations. According to Scheerens et al (2005), schools will want to protect themselves from misunderstanding and outside criticism. Particularly when the stakes are high in external evaluations, schools will tend to present a more positive picture in self evaluations of their functioning (Janssens and Van Amelsfoort, 2008). The fact that self-evaluations are imposed on schools will contribute to self-evaluations being seen more as an obligation to meeting statutory and regulatory obligations, rather than as a tool for improving the school's functioning as an educational institution (Vanhoof, Van Petegem, De Maeyer, 2008). Matthews and Sammons (2004) for example found schools to produce two self evaluations: one for internal use and one for accountability purposes. Accountability officers such as school inspectors therefore have to be very careful in using self-evaluations in assessing schools. Uncritically using results of self-evaluations of schools in external accountability may lead to unreliable and invalid assessments of schools. These examples show that the accountability and the improvement function of both internal self-evaluations and external accountability may be hard to combine.

A right balance between internal self evaluations and external accountability is in any case important. As described in this section, it is generally assumed that using internal self-evaluations in accountability systems is effective in terms of improving schools and ultimately student achievement. According to Janssens and Van Amelsfoort (2008), this assumption has so far not been studied. This study therefore provides an overview of different types of accountability systems using internal self-evaluations and analyses the effectiveness of using internal self-evaluations in accountability systems. The following research questions will be addressed:

- How are self-evaluations used in accountability systems?
- In what types of accountability systems are self-evaluations used most?
- How effective are accountability systems using self-evaluations in terms of improving student achievement?

### **Theoretical framework**

First, a theoretical framework is described that may be used to study the use of self-evaluations in external accountability. This framework provides an overview of how self-evaluations of schools are used in external accountability, and the types of external accountability systems in which self-evaluations may be used.

#### *Use of self- evaluations of schools in external accountability*

A number of studies on internal and external evaluation and accountability show how accountability systems differ in the requirements to self-evaluations of schools and the prescription of criteria schools have to use. Studies of Eurydice (2004) and CPRE (2001) for example show that some accountability systems oblige schools to have internal self-evaluations every year or every other year. In these accountability systems schools are often also obliged to use predetermined standards and criteria in evaluating their education; these standards generally match the standards used in external accountability. The better the content of a self-evaluation corresponds with the frameworks used for external accountability, the better self-evaluations are used in external accountability to assess quality of schools.

Kyriakides and Campbell (2004) identify three ways to connect school self-evaluation with external evaluations. Self-evaluations and external evaluations exist parallel when both the school and the external body conduct their own evaluations, and maybe compare and share findings afterwards. This approach is used to overcome some of the side effects of integrating self evaluations in external accountability. External evaluations is used as a means for measuring school effectiveness and reporting reliable results to stakeholders, whereas the results of self evaluations are used by the school in order to improve its effectiveness. The second way to integrate self evaluations and external evaluations is sequential evaluation. According to Kyriakides and Campbell (2004), schools conduct their own evaluation and the external body (for example an Inspectorate) then uses the results of the self evaluation as a basis for its external evaluation. This approach may also work in the opposite direction when the external body provides feedback to the school which is expected to be used in self-evaluations and improvement of the school. In this case however, the action plan for improvement of the school will reflect the criteria and judgments of the external inspectors instead of the school-defined criteria. In the last 'cooperative model' to integration, the parties that are involved in the two forms of school evaluation discuss and negotiate the process of each form of evaluation. As a result, the interests and viewpoints of external evaluators as well as the school-defined criteria are taken into account simultaneously. According to Kyriakides and Campbell (2004), measurement criteria come up from both types of evaluation and an attempt is made to combine the results in order to conduct a holistic evaluation and satisfy the needs of both parties. Optimal correspondence is often also achieved by providing guidelines, instructions and examples to schools (Janssens and Van Amelsfoort, 2008). In some European countries, Inspectorates of Education even provide training in methods of self-evaluation or add external evaluators to the self-evaluation team in the school.

Results of internal self-evaluations may be used to hold teachers, school and/or districts or local communities to account. School inspectors may for example use performance evaluations of teachers conducted by their principals as part of the human resource development of the school to assess individual teachers. Self-evaluations of schools are part of proportional inspections of schools in the Netherlands and in the United Kingdom. States in the U.S. require districts to provide information on the functioning of all the schools in the district to assess performance of that district.

#### *Types of external accountability systems*

The use of internal self-evaluations may be different for different types of external accountability systems. Different countries choose different roles and functions of school self-evaluation in external accountability (Scheerens, van Amelsfoort, Donoughue, 1999). This section describes different types of external accountability and how these systems are expected to use self-evaluations of schools.

The two predominant types of educational accountability are test-based accountability where the results of achievement tests are the main source of information for evaluating school performance, and school inspections where inspectors visit schools to assess educational quality. School inspectors gather information on the functioning of a school via interviews, document study and inspection visits to schools. School self-evaluations may be used as part of proportional inspections. Most Inspectorates of Education assess several aspects of educational processes in, and output of schools, such as the teaching-learning process, the school organizational conditions (such as the school's educational policy, management of financial resources) and student achievement. School inspections generally include observation of lessons, interviews with principals, teachers and sometimes with parents and pupils, questionnaires to be completed by school staff in preparation for the school visits, and the study of school documents (Standaert, 2000). School inspections are referred to as *process-based accountability* systems as they are primarily used to evaluate educational processes in schools. Test-based accountability on the other hand (only) includes output or outcome standards on the results achieved by schools (Eurydice, 2004). Output standards relate to the skills or knowledge that students acquire while at school. 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). These types of accountability systems are named 'results driven accountability'; 'test-based accountability' or '*output-based accountability*'. Educators are accountable for student learning and test scores of students. Output-based accountability generally requires schools to evaluate their output with respect to external standards and targets and provide the accountability agency with self evaluation reports that include student achievement results.

Accountability systems also differ in the extent to which they include incentives that impose high or low stakes for schools. High stakes accountability systems include severe consequences for schools when failing to meet some performance target, whereas low stakes accountability systems provide relatively mild incentives to perform well. 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. Both *high and low stakes accountability systems* may use self-evaluations of schools to hold schools to account.

A distinction is often also made in accountability systems holding schools to account for their performance (justification) or accountability systems aiming to build capacity in schools and to help schools to improve. Contradictions between accountability and improvement-oriented systems were already described by Ewell in 1987. He described how accountability oriented systems use assessments that severely constrain an institution's ability to develop local assessment mechanisms that meet institution's own needs. Aperi et al (1990) also point to the dichotomous perspective of accountability versus improvement, particularly with respect to the nature and direction of using assessments in accountability systems. The accountability-driven approach to using assessments in accountability systems is oriented towards documenting and reporting the strengths and weaknesses of an institution and its programs, using summaries of easily obtainable, quantitative data. The agenda and time tables of these types of accountability are primarily externally driven and institutions are primarily concerned with satisfying external agencies and meeting minimal external standards and targets. Improvement-driven accountability on the other hand is also externally driven, but accepted internally in institutions as a result of joint activities and responsibilities in developing the time table that is an integral part of existing evaluation activities. The emphasis of an improvement-driven

approach is on institutional self-evaluation where assessments are used formatively as part of a feedback loop. The improvement-driven approach to assessment places emphasis on formative processes of review rather than on providing data for summative purposes. Student outcomes are used to improve instructional processes and student experiences. Improvement-oriented accountability systems are therefore expected to make more use of self evaluations of schools to hold schools to account.

The assumption will be explored that accountability systems with an increasing degree of combining self-evaluations with external accountability are most effective. Process-based, improvement-oriented accountability systems are expected to make the most use of internal self-evaluations, and are therefore expected to be most effective in terms of improving student achievement.

## **Methods**

In the study presented here, the results of studies by Eurydice on educational evaluations administered in European countries in 2000/2001, the SICI blue book containing a description of Inspectorates of Education in Europe in 1999 and the database developed by the Consortium for Policy Research in Education on accountability systems in the US (2000) were used to capture data on the use of internal self-evaluations in accountability systems in both Europe and the U.S. These reports include information on the accountability and external evaluation systems in all (50) states in the U.S. and 30 countries in Europe. The state profiles of the Consortium for Policy Research in Education on accountability systems in the US (2000) for example describe the performance indicators schools have to use to provide information. The Eurydice report gives an overview of the approaches to internal and external evaluation of schools providing compulsory education in thirty European countries.

A database was composed describing the following variables:

### *Use of self-evaluations in external accountability*

- Existence of compulsory/recommended internal self-evaluations: the extent to which accountability systems oblige schools to perform self-evaluations as part of external accountability; ranging from no compulsory self-evaluation (0) to recommended self-evaluations (1), variability in recommended and compulsory self-evaluations (2) and compulsory self-evaluations (3).
- The relationship between self-evaluations and external accountability; ranging from no relationship (0), self-evaluations and external accountability are complementary in goals and focus and are conducted parallel (1), external accountability is based partly on the results of self-evaluations (2), to external accountability is based entirely on the results of self-evaluations (3).
- The extent to which self-evaluation criteria are prescribed; ranging from no prescriptions (0), to prescriptions with respect to use of documents (1), recommendations on criteria (2), criteria on output (3) criteria on educational processes (4), to criteria on both output and educational processes (5).
- Number of self-evaluations every year to hold teachers, schools or districts/local communities to account; ranging from no use of self-evaluations of teachers, schools or districts/local communities (0) to using self-evaluations once a year (1).

### *Types of external accountability systems*

- Output-based accountability system: standards, measurement methods (tests), targets and stakes aimed at output of students, teachers, schools, local communities or districts.
- Process-based accountability system: standards, measurement methods (inspection visits), targets and stakes aimed at educational processes on teacher, school, local community or district level.
- High/low stakes accountability: sanctions, rewards and interventions.
- Accountability systems aiming at accountability: externally defined standards, reporting of results, sanctions
- Accountability systems aiming at capacity building and improvement: evaluation of educational processes, visits to schools, offering advice and support.

Scores were given to measurement methods and stakes on a scale from 0-5; the scales on standards and targets include the total number of standards and targets. A higher score on the scale represents more use of measurement methods and higher stakes.

Effects of the use of internal self-evaluations in external accountability on student achievement are measured by analyzing gains in student achievement from grade 4 in 2003 to grade 8 in 2008 and by analyzing differences in performance levels of student achievement in grade 4 and grade 8 between countries and states. 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 (and their use of self-evaluations of schools) 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.

NAEP and TIMSS mathematics test scores are used to determine student achievement in every country and state. These measures test students in similar topics and item format in mathematics in both grades 4 and 8 every four years. NAEP is measured in every state in the U.S., while TIMSS is administered in most countries in Europe. In some years students responded to both NAEP and TIMSS, enabling a direct comparison of performance on the two instruments. As 1999/2000 is the reference year for analyzing accountability systems, 2003 is the first year that can be used to analyze effects of using self-evaluations in accountability systems on student achievement. 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 in measuring gains in student achievement. 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 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. 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) did not yield reliable results (Johnson et al, 2003).

Descriptive information on every variable was analyzed to gain insight into how self-evaluations of schools are used in accountability systems. Correlation analysis was calculated to see in what types of accountability systems self-evaluations are used most. Using regression analysis, the hypothesis is tested that accountability systems using self-evaluations make more rapid gains in student achievement and generally perform at higher levels 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. Also, a correction is made for prior levels of student achievement in 1999/2000.

## **Results**

This section first describes how self-evaluations of schools are used in accountability systems, and in which types of accountability systems they are used most. Next, the results of the regression analyses are presented, explaining effectiveness of accountability systems using results of self-evaluations.

### *Use of self-evaluations in accountability systems*

First, a reliability analysis was used to compose one scale to describe the use of self-evaluations in external accountability. The reliability of the scale is presented in table 1, as well as the item-total correlation of every item in the scale. The internal consistency of the set of items in a scale is generally found to be sufficient when the coefficient has a minimum value of 0.70. The scale shows a reliable alpha of 0.784. The reliability of the scale can be increased by excluding the items on using self-evaluations to hold teachers and districts or local communities to account. Such a scale would however only describe the use of self-evaluations to hold schools to account. The aim of this study is to include all types of accountability systems using self-evaluations. As the overall reliability of the scale (including these items) is sufficient, these items were therefore included in the scale.

Table 1. Reliability analysis on ‘use of self-evaluations in accountability systems’

	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Existence of compulsory/ recommended self-evaluations	0.830	0.658
Relationship between self-evaluations and external evaluations	0.790	0.679
Extent to which self-evaluation criteria are prescribed	0.785	0.701
Number of self-evaluations every year to hold teachers to account	0.233	0.808
Number of self-evaluations every year to hold schools to account	0.745	0.752
Number of self-evaluations every year to hold districts/local communities to account	0.135	0.812
Cronbach’s alpha scale ‘use of self-evaluations in accountability systems’ (6 items)		0.784

Table 2 provides descriptive information on the use of internal self-evaluations in external accountability systems. The results show that a majority of accountability systems in Europe and the U.S. (66.3%) recommend or oblige schools to perform internal self-evaluations as part of external accountability. The results of these internal self-evaluations are used in 57.5% of all accountability systems in Europe and the U.S. They are mostly used to hold schools to account (46.3%), sometimes also to hold districts or local communities to account (16.3%). Results of internal self-evaluations are almost never used to hold teachers to account (8.8%). The criteria schools should use in internal self-evaluations are prescribed in a minority of accountability systems (31.3%). These criteria include a prescription of the documents schools should deliver (13.8%), output-related criteria schools should use (16.3%) or prescriptions on both output and process-related criteria (11.3%). Only a small number of accountability systems prescribe only process-related criteria (3.8%).

Table 2. Descriptive information on use of self-evaluations in accountability systems

	Number and percentage of accountability systems
Compulsory/ recommended self-evaluations in accountability system	53 (66.3%)
Self-evaluations are used in accountability systems	46 (57.5%)
Self-evaluation criteria are prescribed:	25 (31.3%)
- Prescriptions with respect to types of documents to deliver or general matters to address in self-evaluations	11 (13.8%)
- Recommendations on indicators to use in self-evaluations	10 (12.5%)
- Prescriptions on output-related criteria to use in self-evaluations	13 (16.3%)
- Prescriptions on process-related criteria to use in self-evaluations	3 (3.8%)
- Prescriptions on both output and process-related criteria to use in self-evaluations	9 (11.3%)

Self-evaluations are used for accountability of teachers	7 (8.8%)
Self-evaluations are used for accountability of schools	37 (46.3%)
Self-evaluations are used for accountability of districts/local communities	13 (16.3%)

#### *Types of accountability systems using self-evaluations*

The second part of the study addresses the types of accountability systems making most use of self-evaluations. First, a number of scales were composed to describe different types of accountability systems. A distinction was made in output-based, process-based, high/low stakes accountability systems, and accountability systems holding educators to account versus aiming to improve schools. Table 3 shows the alphas of each scale. The scales describing process-based accountability systems and accountability systems aiming to improve schools have a relatively low alpha. Given the relative few number of items in these scales, the alphas are considered to be acceptable and the scales are included in further analyses.

Table 3. Reliability analysis on different types of accountability systems

	Cronbach's Alpha, imputed database	Number of items
<i>Output-based versus process-based accountability systems</i>		
Output-based accountability systems (Standards, measurement methods (tests), aimed at output)	0.771	5
Process-based accountability systems (Standards, measurement methods (inspection visits) aimed at educational processes)	0.625	8
<i>High stakes versus low stakes accountability</i> (Sanctions, rewards and interventions)	0.736	6
<i>Goals of accountability systems</i>		
Accountability systems aiming at accountability (Externally defined standards, reporting of results, sanctions)	0.677	5
Accountability systems aiming at capacity building and improvement (Evaluation of educational processes, visits to schools, offering advice and support)	0.605	6

A correlation analysis was calculated to provide insight in the types of accountability systems using self-evaluations of schools. Table 4 shows that process-based, low stakes accountability systems aiming to improve schools mostly include self-evaluations of schools in evaluating teachers, schools or districts/local communities. Output-based, high stakes accountability systems that aim to hold educators to account make less use of self-evaluations of schools.

Table 4. Correlations between types of accountability systems and use of self-evaluations

	Use of self-evaluations in accountability systems
Output-based accountability systems	-0.258(*)
Process-based accountability systems	0.424(**)
High stakes versus low stakes accountability	-0.437(**)
Accountability systems aiming at accountability	-0.452(**)
Accountability systems aiming at capacity building and improvement	0.352(**)

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



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

*Effectiveness of using self-evaluations of schools in accountability systems*

The last part of this study includes the effectiveness of using self-evaluations in accountability systems. In a regression analysis the hypothesis is tested that using self-evaluations in external accountability contributes to countries and states making more rapid gains in student achievement and generally performing at higher levels of student achievement. Prior achievement, educational expenditure and pupil/teacher ratios are added to the regression analysis to correct for differences in circumstances and policies of each state and country.

Effectiveness of the use of self-evaluations in accountability systems is measured by relating this independent variable to gains in student achievement and to performance levels in student achievement in grade 4 and 8 in 2003 and 2007. The results in table 5 show that accountability systems using self-evaluations are not effective. No significant relations were found.

Table 5. Effectiveness of using self-evaluations of schools in accountability systems

	Dependent variables				
	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
Independent variables:					
Average student achievement in mathematics grade 8 1999/2000	0.079 (0.053)	0.457** (0.085)	0.641** (0.064)	0.381** (0.098)	0.536** (0.082)
Percentage of GDP on public (pre)primary, secondary and post secondary education	3.405 (1.721)	-2.757 (2.775)	-1.584 (2.084)	3.571 (3.199)	0.648 (2.677)
Pupil/teacher ratio for pre-primary, primary and secondary education	-0.466 (0.468)	-0.022 (0.756)	-0.327 (0.567)	0.476 (0.871)	-0.488 (0.729)
Use of self-evaluations in accountability systems	-0.328 (0.876)	-1.280 (1.413)	-0.764 (1.061)	-1.822 (1.629)	-1.607 (1.363)
Constant	-55.209 (30.778)	296.036** (49.640)	191.867** (37.270)	311.474** (57.224)	240.828** (47.884)
R square	0.100	0.303	0.595	0.175	0.385
Sample size	57	60	65	64	60

Note: standard error in parentheses, \* p < .05, \*\* p < .01

Next, a regression analysis was performed for the different types of using self-evaluations in external accountability to see if specific aspects of using self-evaluations are effective. Table 6 presents the results of the regression analysis. The results show that the existence of compulsory self-evaluations in external accountability has a negative effect on performance levels of student achievement. The other types of using self-evaluations in external accountability have no effect on student achievement.

Table 6. Effectiveness of different types of use of self-evaluations of schools in accountability systems

	Dependent variables				
	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
Independent variables:					
Average student achievement in mathematics grade 8 1999/2000	0.243* (0.093)	0.644** (0.146)	0.889** (0.119)	0.355* (0.164)	0.789** (0.133)
Percentage of GDP on public (pre)primary, secondary and post secondary education	1.708 (2.740)	1.351 (4.283)	4.463 (3.503)	3.882 (4.818)	3.940 (3.920)
Pupil/teacher ratio for pre-primary, primary and secondary education	0.062 (0.744)	-1.161 (1.164)	-0.036 (0.952)	-1.144 (1.309)	-0.660 (1.065)
Existence of compulsory self-evaluations	0.337 (2.119)	-8.866* (3.311)	-4.286 (2.708)	-9.685* (3.726)	-7.455* (3.031)
Relationship between self-evaluations and external evaluations	-0.535 (2.088)	5.914 (3.264)	1.423 (2.670)	4.968 (3.627)	2.713 (2.988)
Extent to which self-evaluation criteria are prescribed	0.795 (1.203)	-0.443 (1.881)	0.230 (1.538)	1.614 (2.116)	1.291 (1.722)
Number of self-evaluations every year to hold teachers to account	-1.652 (5.453)	9.721 (8.524)	11.348 (6.971)	6.371 (9.590)	10.199 (7.803)
Number of self-evaluations every year to hold schools to	-2.676 (4.485)	12.341 (7.010)	7.954 (5.733)	11.585 (7.887)	8.061 (6.417)

	Dependent variables				
	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
account					
Number of self-evaluations every year to hold districts/local communities to account	-3.975 (4.348)	3.625 (6.797)	0.994 (5.559)	-4.501 (7.647)	0.380 (6.222)
Constant	-138.235* (56.766)	208.253* (88.728)	41.338 (72.565)	354.009** (99.824)	106.736 (81.220)
R square	0.240	0.518	0.671	0.336	0.600
Sample size	55	54	52	64	60

Note: standard error in parentheses, \* p < .05, \*\* p < .01

#### *Explaining effects of using self-evaluations of schools in external accountability*

The negative effect of the use of self-evaluations of schools in external accountability may be explained by the type of accountability system they are part of. The results of the correlation analysis in table 4 show that accountability systems using self-evaluations are mostly used in process-based, improvement-oriented, low stakes accountability systems. The negative effect of using self-evaluations in external accountability may be explained by an overall negative effect of this type of accountability system they are used in. Two regression analyses were used to analyze if this is the case. The first analysis (table 8) includes the use of self-evaluations in process-based, improvement-oriented, low stakes accountability systems. The second analysis in table 9 includes the results of using self-evaluations in output-based, accountability-oriented, high stakes systems. The results show that the negative effect of compulsory self-evaluations in external accountability disappears when the type of accountability system they are part of is included in the analysis. Surprisingly, this negative effect of compulsory self-evaluations disappears when both types of accountability systems are added to the analysis. Instead, process-based, low stakes, improvement-oriented systems seem to have a negative effect on student achievement in grade 4 and 8 in 2007. Accountability-oriented systems also have a negative effect on student achievement in grade 8 in 2007; high stakes accountability however positively contributes to student achievement in grade 8 in 2007.

Table 8. Effectiveness of using self-evaluations in process-based, low stakes improvement-oriented accountability systems

	Dependent variables				
	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
Independent variables:					
Average student achievement in mathematics grade 8 1999/2000	0.272** (0.095)	0.716** (0.167)	0.966** (0.130)	0.507** (0.175)	1.067** (0.117)
Percentage of GDP on public (pre)primary, secondary and post secondary education	1.535 (2.283)	3.334 (3.999)	4.577 (3.118)	5.731 (4.194)	4.566 (2.811)
Pupil/teacher ratio for pre-primary, primary and secondary education	0.254 (0.616)	-0.169 (1.079)	0.466 (0.841)	0.247 (1.131)	0.754 (0.758)
Existence of compulsory self-evaluations	0.143 (1.139)	-0.066 (1.995)	0.985 (1.555)	0.717 (2.092)	1.676 (1.402)
Process-based accountability systems	2.120 (8.880)	9.960 (15.554)	3.792 (12.128)	12.348 (16.312)	-29.949** (10.933)
Low stakes accountability systems	1.056 (2.218)	-7.747 (3.885)	-5.127 (3.030)	-10.671* (4.075)	-6.918* (2.731)
Improvement oriented accountability systems	-5.269 (4.462)	-1.793 (7.815)	-4.258 (6.094)	-5.661 (8.196)	7.647 (5.494)
Constant	-153.959** (56.033)	156.200 (96.396)	0.787 (75.166)	259.437* (101.097)	-53.618 (67.760)
R square	0.262	0.430	0.646	0.317	0.721
Sample size	57	60	65	64	60

Note: standard error in parentheses, \* p < .05, \*\* p < .01

Table 9. Effectiveness of using self-evaluations of schools in output-based, high stakes accountability-oriented systems

	Dependent variables				
	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
Independent variables:					
Average student achievement in mathematics grade 8 1999/2000	0.336** (0.094)	0.653** (0.164)	1.047** (0.129)	0.487** (0.177)	1.113** (0.118)
Percentage of GDP on public (pre)primary, secondary and post secondary education	2.414 (2.217)	2.653 (3.856)	5.277 (3.024)	5.790 (4.172)	6.212* (2.774)
Pupil/teacher ratio for pre-primary, primary and secondary education	0.176 (0.583)	0.058 (1.014)	0.453 (0.795)	0.496 (1.097)	0.501 (0.729)
Existence of compulsory self-evaluations	-0.723 (1.048)	0.793 (1.823)	0.275 (1.430)	0.953 (1.973)	-0.004 (1.312)
Output-based accountability systems	0.131 (-.234)	-0.354 (0.407)	0.233 (0.320)	-0.331 (0.441)	0.322 (0.293)
High stakes accountability systems	20.300* (8.436)	-16.313 (14.673)	22.735 (11.509)	1.387 (15.876)	46.411** (10.556)
Accountability oriented systems	-21.379* (8.212)	25.210 (14.283)	-18.566 (11.204)	10.879 (15.454)	-36.089** (10.275)
Constant	-174.132** (52.853)	153.054 (91.928)	-43.029 (72.105)	232.641* (99.463)	-73.575 (66.131)
R square	0.327	0.460	0.660	0.311	0.730
Sample size	57	60	65	64	60

Note: standard error in parentheses, \*  $p < .05$ , \*\*  $p < .01$

### Conclusion and discussion

This paper describes accountability systems using self-evaluations of schools to assess educational quality in schools. These accountability systems are expected to be most effective as quality assurance and self-evaluations in schools are stimulated through external accountability. Schools are motivated to use external accountability standards to organize their self-evaluations and school boards are expected to use these standards in the governance of their educational processes and functioning of their schools. Until now the validity of these assumptions has not been studied to a large extent. This

study therefore provides an overview of different types of accountability systems using self evaluations and analyses the effectiveness of using self evaluations in accountability systems.

The results of the study show that a majority of accountability systems in Europe and the U.S. recommend or oblige schools to perform self-evaluations as part of external accountability. The results of these self-evaluations are used in a slight majority of all accountability systems in Europe and the U.S. They are mostly used to hold schools to account, sometimes also to hold districts or local communities to account. Results of self-evaluations are almost never used to hold teachers to account. The criteria schools should use in self-evaluations are prescribed in a minority of accountability systems. Process-based, low stakes accountability systems aiming to improve schools mostly include self-evaluations in evaluating teachers, schools or districts/local communities. Output-based, high stakes accountability systems that aim to hold educators to account make less use of self-evaluations of schools.

The use of self-evaluations in external accountability in general does not seem to have an effect on student achievement. When analyzing effectiveness of specific aspects of using self-evaluations in accountability systems, the results even show that the existence of compulsory self-evaluations in external accountability has a negative effect on gains in student achievement and on performance levels of student achievement. This negative effect can be explained by the type of accountability system self-evaluations are part of. The negative effect of compulsory self-evaluations in accountability systems disappears when the type of accountability system it is part of is taken into consideration. Both process-based, improvement-oriented, low stakes accountability systems and output-oriented, high stakes, accountability oriented systems diminish the negative effect of compulsory self-evaluations in external accountability. These results lead to the conclusion that self-evaluations in external accountability overall are not effective, and may even be ineffective. This last conclusion is however presented with caution as the negative effect is rather small and may also result from measurement bias that may have occurred when linking TIMSS and NAEP test scores.

As convincing positive effects of using self-evaluations of schools in external accountability are however lacking, we ask ourselves how to explain this conclusion. A potential explanation may be that using self-evaluations in accountability may lead to goal displacement and side effects such as window dressing or tunnel vision. Goal displacement is described by Warner and Havens (1968) as neglectance of major organizational goals in favor of goals associated with building or maintaining the organization. An important reason for the displacement of goals is the intangibility of organizational goals such as goals on educational quality. Using self-evaluations in external accountability may lead to goal displacement when quality controls of schools focus on checking procedures and documents of schools on, for example, quality assurance, instead of assessing how self-evaluations are used to assess and improve how well teachers and school leaders are working towards high student achievement. External evaluators often only check if schools have some procedure and protocol in place for self-evaluation, instead of assessing the quality of these procedures. This may lead to a number of side effects. Schools may for example focus on implementing procedures and protocols that have no (positive) effect on primary processes in the school, but are implemented to be assessed more positively (window dressing), or schools may emphasize phenomena that are quantified in the standards of the accountability system (such as the number of lesson hours) at the expense of unquantifiable aspects of performance, such as the quality of lessons (tunnel vision) (De Wolf and Janssens, 2005; Smith, 1995). Janssens and Van Amelsfoort (2008) also give examples of schools writing self evaluations for the Inspectorate which no longer serve the goals of improving education but of providing a description and analysis of how well the school is doing.

Also, the results of self-evaluations are often used by external evaluators to judge educational quality and student achievement in schools without checking the validity and reliability of these results. The information on student achievement, delivered to the states in the U.S. by individual schools, is for example in most cases published on the internet without any check of the accuracy of these scores, leaving schools the opportunity to manipulate their test scores to be evaluated more positively. Jacob and Levitt (2003) for example found 4-5% of teachers cheating on standardized tests used in external accountability to improve test scores. This type of fraud is less likely to happen when tests are only part of self-evaluations or when test scores are checked by external evaluators. The intended

mechanism of accountability systems stimulating schools to improve achievement of their students is jeopardized.

These side effects may explain why the use of self-evaluations in these systems has a negative effect on student achievement. A potential recommendation to using self-evaluations in external accountability may therefore be to use an approach suggested by Janssens and Van Amelsfoort (2008). They opt for a participatory approach in which evaluation criteria are drawn up by the school on an entirely autonomous basis, though naturally with due regard given to local and national objectives. External evaluators, for their part, should take the evaluation criteria and objectives drawn up by schools during self-evaluation into account, so that they are able, if necessary, to qualify their own judgments. Janssens and Van Amelsfoort (2008) expect this kind of open inspection framework that leaves space for the schools' own priorities and audit trails, to be a way to combine the advantages and disadvantages of using the inspection framework for purposes of self-evaluations.

Another recommendation may be to carefully design criteria to judge the value of self-evaluations with respect to providing information on the functioning and improvement of schools. External evaluators should only use the results of self-evaluations if these evaluations meet these criteria. In any case, more research is needed to find out how self-evaluations can best be used in external accountability systems.

## References

- Aper, J.P., Cuver, S.M., Hinkle, D.E. (1990). Coming to terms with the accountability versus improvement debate in assessment. *Higher Education*, 20, 471-483.
- Barber, M. (2004). The Virtue of Accountability: system Redesign, Inspection, and Incentives in the Era of Informed Professionalism. *Journal of Education*, 85(1), 7-38.
- Booher-Jennings, J. (2005). Below the Bubble: 'Educational Triage' and the Texas Accountability System. *American Educational Research Journal*, 42(2), 231-268.
- 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](http://www.cpre.org/index.php?option=com_content&task=view&id=169&Itemid=124)
- Cousins, J.B. and Whitmore, E. (1998). Framing Participatory Evaluation. *New Directions for Evaluation*, 80, winter, 5-23.
- Eurydice (2004). Evaluation of Schools providing Compulsory Education in Europe. <http://www.eurydice.org/portal/page/portal/Eurydice>
- Ewell, P.T. (1987). Assessment, accountability and Improvement; Managing the Contradiction. *Paper presented for the American Association for Higher Education Forum*.
- Goertz, M.E. and Massell, D. Summary. In: B. Gross and M.E. Goertz (2005). *Holding High Hopes: How High Schools Respond to State Accountability Policies*. CPRE Research Report Series, RR-056.
- Hanushek, E.A. and Raymond, M.E. (2001). The Confusing World of Educational Accountability. *National Tax Journal*, 54(2), 365-384.
- 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.
- Janssens, F.J.G. (2005). *Toezicht en Governance*. Universiteit Twente: oratie.
- Janssens, F.J.G. and Van Amelsvoort, G.H.W.C.H. (2008). School self-evaluations and school inspections in Europe: An exploratory study. *Studies in Educational Evaluation* 34 (2008), 15–23.
- 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.M. (2002). Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *The Journal of Human Resources* (37(4), 752-777.
- Kyriakides, L. and Campbell, R.J. (2004). School Self-evaluation and School Improvement: A Critique of Values and Procedures. *Studies in Educational Evaluation*, (3), 23-36.
- Matthews, P. & Sammons, P. (2004). *Improvement through Inspection*. London: Ofsted.
- 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](http://nces.ed.gov/timss/pdf/naep_timss_pisa_comp.pdf)
- Nevo, D. (2001). School evaluation: internal or external? *Studies in Educational Evaluation*, 27, 95-106.



SICI (1999). *Inspectorates of Education in Europe; A descriptive study*. Brussel: DVO

Scheerens, J., Van Amelsfoort, G. and Donoughue, C. (1999). Aspects of the organizational and political context of school evaluation in four European countries. *Studies in Educational Evaluation*, (25), 79-108.

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.

Vanhoof, J., Van Petegem, P. and De Maeyer, S. (2008). Attitudes Towards School Self Evaluation. Linking the Individual View to the School Context. *Paper presented at ECER 2008*.

Vanhoof, J., Petegem, P. van (2009). The Process and Result of School Self Evaluations. How Different are the Quality Judgments of School Principals and Inspectors? *Paper presented at EARLI*

Warner, W.K. and Havens, A.E. (1968). Goal Displacement and the Intangibility of Organizational Goals. *Administrative Science Quarterly*, 12(4), 539-555.

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>