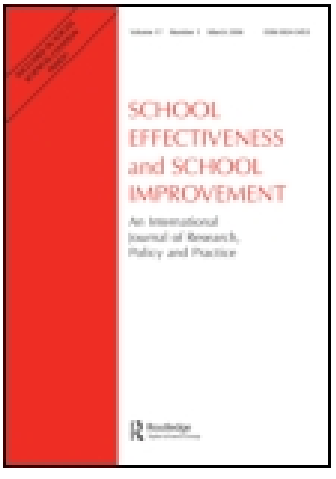

School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nses20

Does accountability pressure through school inspections promote school improvement?

Herbert Altrichter[a] & David Kemethofer[a]

[a] Department of Education and Psychology, Johannes Kepler University, Linz, Austria

Published online: 21 Jan 2015.

To cite this article: Herbert Altrichter & David Kemethofer (2015) Does accountability pressure through school inspections promote school improvement?, School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 26:1, 32-56, DOI: 10.1080/09243453.2014.927369

To link to this article: http://dx.doi.org/10.1080/09243453.2014.927369

School Effectiveness and School Improvement, 2015
Vol. 26, No. 1, 32–56, http://dx.doi.org/10.1080/09243453.2014.927369

Routledge
Taylor & Francis Group

Does accountability pressure through school inspections promote school improvement?

Herbert Altrichter* and David Kemethofer

Department of Education and Psychology, Johannes Kepler University, Linz, Austria

"New" school inspections are essential parts of "evidence-based governance" concepts and have been implemented by many European countries as a major strategy to assure and improve the effectiveness and quality of their education systems. However, national inspection systems vary in their composition and in their contextual features. Using online survey data from approximately 2300 principals in 7 European countries, the paper explores the role of "accountability pressure" as an element for understanding the operation of inspection systems. The results indicate that principals who feel more "accountability pressure" are more attentive to the quality expectations communicated by inspections, more sensitive to stakeholders' reactions to inspection results, and more active with respect to improvement activities. However, also the number of unintended consequences is increasing with pressure. Inspection systems in different countries are seen by school leaders as applying differential degrees of "accountability pressure", which is reflected in system-specific amounts of improvement activities.

Keywords: school inspection; accountability; accountability pressure; comparative research; school development; school improvement

"New" school inspections and the mechanisms of school improvement

Some European Inspectorates of Education (e.g., the Office for Standards in Education, Children's Services and Skills [Ofsted] in England; the Irish, Czech, and Austrian Inspectorates; see Greger, 2011; Scheipl & Seel, 1985) have existed since the mid-19th century. Also the Dutch Inspectorate of Education has been in operation for more than 200 years; however, its working methods have evolved greatly over these years. The Swedish school inspection, which also originated from the 19th century, lost its function to control individual schools in the 1990s (Lindgren, Hult, Segerholm, & Rönnberg, 2012). In 2003, external control of school quality and legal appropriateness was strengthened, and the school inspectorate was again charged with evaluating individual schools (Gustafsson & Myrberg, 2011). In Germany, school inspections have only recently been introduced in response to increased decentralization of education policy or as a reaction to international student assessment results (Kotthoff & Böttcher, 2010).

Clearly, not everything which is labeled "inspection" is operating in identical ways. The traditional inspection role in centralist-bureaucratic states (such as Austria or the Czech Republic) was based on an intermediate position in the hierarchical line located above individual schools but beneath the central administration. Several supervisory functions for schools were amalgamated in this role. Headpersons of schools were directly accountable to these "old inspectors", who were also responsible both for distributing staff

*Corresponding author. Email: herbert.altrichter@jku.at

to individual schools and for "inspecting" and controlling the quality of the schools in their region. Criteria for assessing this "quality" may have varied in time but have always included some mixture between educational effectiveness and quality, on the one hand, and legal and administrative appropriateness, on the other.

In a climate of globalization and international comparisons, evaluation and accountability have become educational key issues in all developed countries (Eurydice, 2004; Maag Merki, 2011). As a consequence, many education systems have introduced varieties of *evidence-based governance regimes*. A "new" type of school inspections is one of the major incarnations of this governance concept. In line with the logic of "evidence-based governance", Inspectorates of Education (a) *set expectations* through their inspection standards and procedures; they (b) *collect evidence* by inspection visits and use information produced by other evaluation instruments to assess the quality of education and hold schools accountable for a broad range of goals related to student achievement, teaching, organization, and leadership; and they (c) aim to *stimulate school and system improvement* by producing reports which point to strengths and weaknesses of individuals schools and include or imply recommendations for action to be undertaken by the inspected schools or the authorities in charge of them (see Ehren, Altrichter, McNamara, & O'Hara, 2013). Thus, inspections mirror the national policy of education and are meant to be an essential system-level factor assuring and promoting the effectiveness and the quality of an education system (see Creemers & Kyriakides, 2012).

What we call "new" inspection systems have been institutionalized in different countries at different times during the last decades; however, they include some recurring features which distinguish them from their predecessors:

(1) "New inspection systems" concentrate on the evaluative functions and shed administrative supervisory functions; consequently, they try to gain some distance or independence from the normal operation of the administrative hierarchy by building up new institutions or emphasizing an "independent expert role" *(differentiation of administrative and evaluative functions)*.
(2) They aim to professionalize these evaluative functions by formalizing and proceduralizing them and by enriching their operation by the use of instruments taken from social science *(expertise through social science)*.

By virtue of these features, "new inspection" fits well in the image of rationalized control propagated by the proponents of evidence-based governance. "New inspectors" are not grey bureaucrats who are supervising schools according to old-fashioned and easily controllable administrative criteria; rather, they aspire to form a new breed of professionals in close contact to the latest developments in educational research.

Compared to the other major instrument of "evidence-based governance" – to performance standards and external performance tests – new inspections have *potential advantages:*

- Inspections usually include output measures but are not restricted to them and may also attend to *process and context information*.
- Data collection and dissemination of results and judgments are more *localized* (e.g., more attention to specific local conditions in inspections and follow-up visits), which offers the opportunity to account for specific local circumstances. This may be conducive to acceptance by schools and their local constituencies.

- They may also provide some *political advantages* in that they may be more flexibly oriented towards new agendas and, thus, make "governments look active" (Hughes, Mears, & Winch, 1997, p. 311).

Comparison with performance standards and external performance tests, however, also points to *potential weaknesses:*

- Data collection and judgments may make inspections more comprehensive and holistic. However, thereby there are also *more judgmental and subjective elements* included which make it easier to challenge inspection results.
- Connected with its localized nature, inspections are *comparatively labor intensive* (both for the inspection providers and schools).

During the last 2 decades, new inspections have been widely implemented by many European countries as one of their major mechanisms to assure and promote the effectiveness and quality of their schools. Educational research has turned to this phenomenon, too. In the meantime, literature reviews have been produced which try to organize and evaluate the findings (see, e.g., De Wolf & Janssens, 2007; Ehren et al., 2013; Ehren & Visscher, 2008; Husfeldt, 2011; Klerks, 2013; Kotthoff & Böttcher, 2010; Luginbuhl, Webbink, & De Wolf, 2009). Husfeldt (2011) distinguishes three types of studies:

(1) *Descriptive studies about attitudes and expectations with respect to inspections:* They frequently observe positive and respectful relationships between those evaluated and those evaluating (e.g., Dedering & Müller, 2011; McCrone et al., 2007); however, they also report on high pressure on the professionals working in schools. Diverging ideas about how inspections will be effective for productive school development can be found in different stakeholder groups; parents and administrators usually expect more effects than staff (Husfeldt, 2011).
(2) *Descriptive studies about reactions to inspections:* Such studies use qualitative and/or quantitative methods to analyze what measures schools and their stakeholders take in the wake of inspections. For instance, Ehren and Visscher (2008) found that schools had developed improvement programs as a consequence of inspections. Matthews and Sammons (2005) recorded a sharp decrease of poor lessons in schools which were in "special measures" after an inspection had unveiled unsatisfactory practices, while reactions were more inconsistent in other schools.
(3) *Studies about improvement of student performance after inspections:* Research of this type originates mostly from the Anglo-American countries. For instance, Cullingford and Daniels (1999), Shaw, Newton, Aitkin, and Darnell (2003), and Rosenthal (2004) did not find overall positive effects of Ofsted inspections on General Certificate of Secondary Education (GCSE) performance, while the studies by Hanushek and Raymond (2005) and Luginbuhl et al. (2009) seem to indicate that inspections can improve student performance. Studies by Shaw et al. (2003) and Matthews and Sammons (2004) point to differential effects for weak, average, and strong schools. Klerks (2013) has produced a systematic meta-review of high-quality peer-reviewed articles published after 2000. She found both positive and negative effects of inspections on student achievement results, which, however, were in either case quite small.

In sum, it seems fair to say that the overall results of inspection research are, at present, far from conclusive as to the question whether or not inspection systems contribute to the aspired quality goals. Husfeldt (2011) has argued that an important reason for inconclusive research findings is to be found in the lack of theoretical models which account for the specific features of inspection approaches and for the in-school processes which mediate between school inspections and their intended mid- and long-term results, such as school's enhanced improvement capacity, high-quality learning conditions, and ultimately improvement of student learning.

Ehren et al. (2013) have recently proposed a conceptual framework for understanding inspection processes, contexts, and results which may remedy this deficit. They reconstructed the processes and mechanisms by which six European inspectorates aim to monitor school quality and stimulate school improvement from an analysis of legal and administrative documents and interviews with relevant officials. Since country-specific results turned out to include wide consensus about major goals and processes between countries, findings were consolidated in a (supranational) "conceptual model" of school inspection. This model identifies three overarching "effective mechanisms" attributed to school inspections (printed in bold in Figure 1). By "*Setting expectations*" and by "*Giving feedback*" (if feedback is accepted and understood by schools), inspections aim to stimulate and drive improvement and self-evaluative actions by the school. The school's *stakeholders* are also expected to *react to inspection standards and reports* and provide additional stimulation to the school to improve. These improvement actions will result in the school's high improvement capacity and in effective school and teaching conditions, which will lead to good student results (Ehren et al., 2013).

But what is the driving and dynamizing element which fuels these processes? What makes school leaders attend to the expectations conveyed by inspection standards and procedures, to the feedback given by inspection reports, and to stakeholders' views and actions instead of doing nothing or other things? "New inspection systems" have been introduced in many European countries as a measure of system control. In a situation of globalized national competition, they were to provide both rational evidence about effective starting points for improving school effectiveness and "systemic pressure" to actually go ahead with it (Altrichter & Heinrich, 2007).

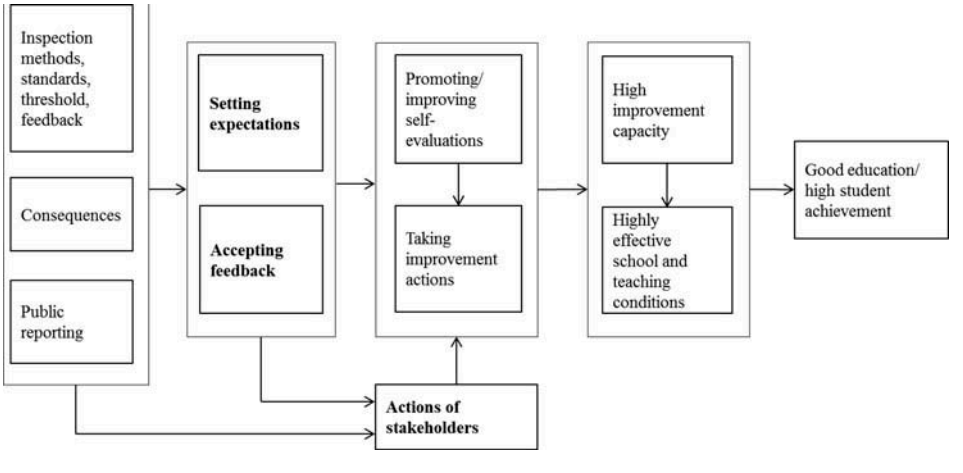

Figure 1. Conceptual model of school inspections (Ehren et al., 2013, p. 14).

"*Pressure to improve*" may be a relevant driving force for improvement by inspections. It is cited by Reezigt and Creemers (2005; see, also, Creemers, Stoll, & Reezigt, 2007) as one of three external dimensions that are important for effective school development. The other two dimensions named by Reezigt and Creemers (2005) are "resources for improvement" and "educational goals". In another paper (Altrichter & Kemethofer, 2012), we found in the data set which is used here that only "pressure to improve" distinguished between the participating national accountability systems, while (our operationalizations of) "resources for improvement" and "educational goals" made no difference. "Pressure to improve" might be crucial for understanding the operation of inspection systems which purport to be an "external condition for effective school development". Some inspection systems include minimal thresholds, labeling of schools, and sanctions for failing schools based on the assumption that "pressure" on schools (through administrative consequences, stakeholder pressure, and/or competitive advantages) is an important lever to make schools conform to inspection standards and to react on inspection results. In this perspective, inspection criteria and interventions are instruments to "deter" rational actors from deviating from the modes of operations considered legitimate by the regulating authority (Braithwaite, 2008).

There are a number of empirical studies which found that "accountability pressure" promotes development activities of schools. Chiang (2009) found that sanction threats boost school spending on teacher training, curricular development, and instructional technology. In a case study in one English secondary school, Perryman (2010) showed that a number of development activities would not have been possible without external pressure. Kotthoff, Maag Merki, and Böttcher (2007) understand inspection consequences as a feature which helps inspection results to be taken seriously by schools. Van Bruggen (2010) argues that schools need external pressure to transfer feedback into practice. Faubert (2009) points to positive effects which are associated with the public release of school results as publications promote pressure for quality. The positive impact of accountability pressure on student achievement results is emphasized by Hanushek and Raymond (2005) and Lee (2006); however, both studies only identified associations for specific subgroups or effects in combination with additional support.

However, "pressure" may also be detrimental for development. There is some evidence that pressure on schools is also associated with unintended negative effects. Bellmann and Weiß (2009) categorized more than 20 unintended effects on class and school level including narrowing of the curriculum, discouraging new teaching strategies, damaging work satisfaction, and cheating of students and teachers (e.g., Kotthoff, 2003; Maag Merki, 2010; Perryman, 2007). Ehren and Swanborn (2012) analyzed unintended consequences of the Dutch accountability system and found that about 30% of the schools exempted students who were expected to perform poorly from sitting accountability-relevant testing. It has also been shown that less effective schools "did not manage to improve their status simply because of the pressure placed upon them" (Creemers & Kyriakides, 2012, p. 51; see, also, Good, Wiley, & Sabers, 2010; Murphy, 2009).

There may be also different forms of "pressure" which are relevant for school improvement. There may be a generalized "achievement press", an appreciation for scholastic performance in a society, which has been shown to be positively associated with student achievement (Creemers & Kyriakides, 2012; Valverde & Schmidt, 2000). There may be forms of pressure which are connected with task and performance, and other ones which are associated with relationships and social power.

It is beyond the scope of this paper to discuss a wide range of possible meanings of "pressure" for various aspects of school development. In the following sections of the

paper, we will concentrate on investigating the role of "accountability pressure" for the processes and results of inspection systems. *Accountability pressure* is defined as pressure on individual schools and their representatives to act in conformity with the standards of an accountability system and to take action to improve school quality and effectiveness (see, also, Reed, Scull, Slicker, & Winkler, 2012). In the next section, we will address the question whether or not "accountability pressure" is a characteristic which is appropriate to distinguish different national inspection systems.

"Accountability pressure" and inspection approaches

In a globalized world, there are a number of educational policies which seem to "travel" in the sense that they are taken up by various national systems in a similar way (see Ozga & Jones, 2006). However, when they are "embedded" in different national educational traditions, seemingly similar instruments might become implemented in different ways and might fulfill different functions. "New school inspection" is obviously one of these traveling policies which has been taken up by nearly all European education systems.

The "conceptual model" by Ehren et al. (2013) suggests that inspection systems in various European countries are built on similar assumptions about intermediate mechanisms and processes necessary for inspections to be effective. Nevertheless, different national inspection systems do not apply identical inspection measures to stimulate these processes. In the following section, we want to look for relevant variations in inspection systems which might make a difference for the way inspections work and produce effects and also for the amount of "accountability pressure" which emanates from them.

According to Maritzen (2007), pressure on school leaders to enforce school improvement will increase if stakeholders notice and make use of inspection results. For Reed et al. (2012, p. 5), six essential elements can be found in "strong" accountability systems: (a) Clear standards must be defined and assessed rigorously; (b) assessment results including data are to be reported to all stakeholders to ensure continuous improvement. Such evaluations should (c) allow regular assessment of individual schools as well as comparisons with other schools. Finally, rewards and consequences should enforce improvement at the (d) school level, (e) the individual student level, and (f) the individual teacher and administrator level. In the following analysis in Table 1, we follow Ehren et al. (2013), who distinguish four main dimensions of school inspections: (a) Types of inspection, (b) Standards, thresholds, and feedback, (c) Consequences, and (d) Reporting. We will explain these dimensions, give some examples how they are implemented in various countries (based on the information collected in Ehren et al., 2013, pp. 32–40), and discuss which features may put special "accountability pressure" on schools to do well in inspections.

Types of inspection

Inspectorates may use "cyclical" school inspections of every school and "differentiated" inspections which are targeted at particularly weak schools. Both types are often used side by side by the same inspectorate; however, an emphasis on "differentiated" inspections often implies that the balance of resources and expertise shifts from "regular cyclical" to "differentiated" inspections. "Thematic" school inspections do not evaluate individual schools but focus on broad topics of teaching and schooling (e.g., on the state of inclusion in schools, the teaching in specific subject areas, or the

Table 1. Summary of inspections characteristics.

	The Netherlands (NL)	*England (ENG)*	*Sweden (SE)*	*Ireland (IE)*	*Czech Republic (CZ)*	*Austria (AT; Styria)*	*Switzerland (CH)*
Types of inspection							
Cyclical inspections of all schools	Every 4 years	Every 5 years	Every 4–5 years	Every 5 years	Every 3 years	Every 2–4 years	Every 4–5 years
Differentiated inspections	*x*	*x*	*x*	–	–	–	–
Thematic school inspections	x	x	x	x	x	–	x
Standards							
Legal aspects	x	x	x	x	x	x	x
Context and process quality	x	x	x	x	x	x	x
Outcomes	*x*	*x*	–	–	*x*	–	*[x]*[1]
Thresholds for distinguishing failing schools	*x*	*x*	–	–	*x*	–	–
Consequences							
(Advising on) sanctions	*x*	*x*	*x*		*x*	–	–
Interventions	x	x	x	x	x	x	x
Reporting							
General/thematic reports	x	x	x	x	–	–	–
Reports on individual schools to the general public	*x*	*x*	–	*x*	–	–	–
Scores on "pressure scale"	**5**	**5**	**2**	**2**	**3**	**0**	**0 – 1**

Notes: "x" indicates the presence of characteristics/mechanisms in each country. Dimensions printed in italics are assumed to add to "pressure" on individual schools. The hits on these dimensions add up to the score on the "pressure scale" in the last line of the table.

[1]Student outcomes are only used in some Swiss cantons which have introduced standardized testing.

The table was adapted from Ehren et al. (2013, p. 20). Additional information for Switzerland was supplemented by Dr. Guri Skedsmo: In Switzerland, 18 out of 21 German-speaking cantons have implemented external evaluation which is conducted by a cantonal department for external evaluation or school supervision or by a university college assigned by the cantonal education authorities (Huber, 2011; Skedsmo & Huber, 2013).

use of ICT in schools) and give overview reports and recommendations for further development on various system levels.

When it comes to assessing what constellations of inspection elements put what amount of "accountability pressure" on schools, our hypotheses are: The more often cyclical inspections take place, the more pressure is experienced by schools. However, there is not much variation in the frequency of inspection visits between the national inspection systems in our study. All systems plan to inspect all schools every 4 to 5 years. (The original idea in Austria to inspect on a tighter schedule proved too ambitious.)

We assume that "differentiated" inspections put more "accountability pressure" on schools than "cyclical" inspections because they include the possibility of being exposed of being in need of special measures. Alternatively, one might hypothesize that a regime focusing on "differentiated" inspections reduces "accountability pressure" on a majority of schools which are not in danger of being subjected to this inspection type, but increases pressure on schools which are close to or under the threshold level.

We assume that "thematic" inspections are not experienced as putting "pressure" on individual schools.

Standards

Standards clarify the criteria against which schools will be evaluated by inspections. All inspection systems in our study check the operation of the schools for compliance with legal regulations (e.g., equal access for all students, minimum lesson hours, security standards), and they apply some criteria for the quality of the context (e.g., school-specific location and student population) and the processes (e.g., school climate, teacher–student relationship, quality of subject teaching) the individual school offers. Some countries also include student results in national performance tests as an important indicator of student quality.

Some inspection systems include "threshold levels", which are used to award grades to schools that often range from "failing" to "well developed" and to identify "failing" schools which do not reach a minimum score on the inspection standards. Threshold levels are usually used for (public) labeling of schools. Ofsted, for example, grades schools as "outstanding", "good", "satisfactory", or "inadequate". Hanushek and Raymond (2002) assume that "schools that have scores close to a threshold might be expected to alter their behaviour more than schools further away from the established critical thresholds" (p. 18).

We hypothesize that the use of output information on student performance puts more "pressure" on schools than other criteria, because much public attention is focused on the ability of schools to help students to produce performance and to earn certificates, and because output data are usually of high measurement quality and comparative nature. Legal criteria, in contrast, are of long-term nature and usually in easy reach for "normal schools".

We also assume that "threshold levels" are indicators for more pressure in inspection systems because they are connected with public labeling of schools and provide grounds for singling out schools for special measures.

Consequences

Inspections are meant to be consequential: They are to stimulate and orientate school improvement. All inspection systems in our study include some type of "intervention", for example, inspection reports spell out recommendations for improvement, schools are asked to write development plans which address weak points and devise ways of building up strengths, and schools are monitored whether or not to implement improvement plans.

"Sanctions" may include more direct interference in the operation of schools, financial fines, or closure of a school. The inspectorates in our study usually do not have the capacity to directly sanction schools, with the exception of the Swedish Inspectorate, which may withdraw the license and funding of independent schools and may temporarily

close down public schools. The Dutch, English, and Czech inspectorates are in the position to advise their education ministry to impose sanctions on failing schools.

Again, we assume that the capacity of inspectorates to impose (or advise) "sanctions" indicates more "accountability pressure" in a system.

Reporting

All inspection systems aim to stimulate both system and individual school improvement by giving reports on the state of schools and the school system. From the perspective of "accountability pressure" on individual schools, it makes a difference whether or not reports on the quality of individual schools are published to their stakeholders or even to the whole public. Hanushek and Raymond (2005) argue that "reporting results has minimal impact on student performance and that the force of accountability comes from attaching consequences …" (p. 298).

The English, Irish, and Dutch inspectorates publish such reports. In The Netherlands, even lists of failing schools and summaries of the inspection assessments of all schools are handed over to the public. In most of the Swiss cantons and similarly in Austria/Styria, the results are not made publicly accessible. In many Swiss cantons, even municipal and cantonal authorities receive anonymized reports so that they do not know single school results (Skedsmo & Huber, 2013). It is the schools which are given an individualized report; it is their responsibility to inform the school governing bodies, parents, and so forth.

We assume that "accountability pressure" is higher in systems which publish inspection reports on individual schools to the general public.

To sum up this analysis: In Table 1, we have collated information about characteristic features of inspection in seven European countries which are participating in our study. Inspection features printed in italics are considered to contribute to more "accountability pressure" in a national context. In order to arrive at a hypothesis of differential "accountability pressure", we add up the scores of national inspection systems in these dimensions. According to a widely used concept for distinguishing national governance and evaluations arrangements (e.g., Maag Merki, 2010), systems with top scores may be considered "high-stake inspection systems", and those with low scores are "low-stake inspection systems". The bottom line of Table 1 shows that – according to this rather crude procedure – the English and the Dutch inspection systems include more elements which are assumed to contribute to "accountability pressure on individual schools", while the Austrian and the Swiss systems include least of such features. The Swedish, the Irish, and the Czech inspection systems seem to occupy middle grounds. In the subsequent sections, we will check whether or not this result of an external analysis of documents is reflected by the experiences of school leaders.

Research design

Research questions and hypotheses

In the following passages, we will empirically explore the role of "accountability pressure" in processes of inspection. In particular, we want to discuss the following research questions and hypotheses:

(1) Does "accountability pressure" promote development activities and/or unintended effects as a consequence of school inspections?
According to Reezigt and Creemers (2005) and to the self-description of many inspection systems (see Ehren et al., 2013), we assume:

Hypothesis 1 (H 1): School leaders who feel more "accountability pressure" say that more development activities are taken in their schools.

Based on findings that "pressure" may also be detrimental for development (e.g., Kotthoff, 2003; Maag Merki, 2010; Perryman, 2007), we assume:

H2: School leaders who feel more "accountability pressure" say that more unintended effects take place in the wake of school inspections.

(2) Are different national inspection systems characterized by differential amounts of "accountability pressure"?
Based on our discussion in the previous section, which is summarized in Table 1, we assume that national inspection systems may be distinguished by the amount of "accountability pressure" felt by school leaders:

H3: School leaders in "high-stake inspection systems" (such as England) say that they feel more "accountability pressure" than those in "low-stake systems" (such as Austria).

(3) Do national inspection systems which are characterized by more "accountability pressure" promote more, less, or qualitatively different development activities and do they produce more unintended and negative consequences?
As a consequence of the assumptions for Hypotheses 1 and 2 and of the clarification with respect to national inspection systems included in Hypothesis 3, we assume:

H4: School leaders in "high-stake inspection systems" say that more development activities are taken in their schools compared to those in "low-stake inspection systems".
H5: School leaders in "high-stake inspection systems" say that more unintended consequences (such as narrowing the curriculum, discouraging new teaching strategies) are to be observed in the wake of inspections.

(4) How does "accountability pressure" impact on school improvement and the underlying mechanisms of inspection?
No hypothesis is formulated for this research question: In an explorative mode, we will ask if there is a differential impact of "accountability pressure" on the processes of school inspection suggested by Ehren et al. (2013).

Data sources and methods

The data used for discussing these questions originates from the European Union-funded project "Impact of school inspection on teaching and learning" (Ehren et al., 2013).[1] Approximately 2,300 primary and secondary school principals in seven European

countries (Netherlands (NL), England (ENG), Sweden (SE), Ireland (IE), Austria (AT; Region Styria), Czech Republic (CZ), and Switzerland (CH; 5 German-speaking cantons Aargau, Appenzell Ausserhoden, Graubünden, Schwyz, Zug[2]) participated in an online survey to collect comparative data on the mechanisms and processes of school inspections in different accountability systems. The sampling procedure differed between the countries: In smaller inspection systems, all schools were included (Ireland, Austria/Region Styria, Switzerland/5 cantons), while other countries used a random sampling strategy (Sweden, Czech Republic). In England and The Netherlands, the sampling originally followed a regression-discontinuity design on the basis of inspection results. As the response rate initially did not meet the expected figures, the sampling strategy was changed in England to random sampling of a larger target group. In The Netherlands, three groups of schools were part of the target sample: "high-risk", "risk", and "no-risk" schools. As a consequence of oversampling of "risk and high-risk" schools, it cannot be ruled out that schools that experience above-average pressure are overrepresented. Because of this fact, Dutch data will not be included into country-specific analyses and only presented in Tables 6 and 7 for illustrative reasons. Table A1 in Appendix 1 provides a summary of sample size and characteristics in each country.

The questionnaire included 73 questions based on the "conceptual model" presented above. Principals scored items about the amount of time used for (improvement) actions the school had taken before, during, or after school inspections on a 5-point scale ranging from *much less time* (= 1) to *much more time* (= 5). A 5-point scale ranging from *strongly disagree* (= 1) to *strongly agree* (= 5) was used when asking for intermediate processes, effective school conditions, and unintended consequences.

The exclusive use of questionnaire data implies that we have to work with information that reflects the principals' subjective reaction to the items. This mode of data collection is adequate to reflect "accountability pressure", which is conceived a subjective reaction of actors. However, it might not be the best way to collect information about development activities. Additionally, the data source only provides information about principals' views and excludes other actors. Nevertheless, we consider the data source adequate for an explorative discussion of our questions if these limitations are kept in mind.

For measuring "accountability pressure", we used the school leaders' reaction to the item "I feel pressure to do well on the inspection standards" (see Table 2). For discussing Research questions 1 and 2, the whole data were used. Two subgroups were built: All school leaders who agreed or strongly agreed to the item "I feel pressure to do well on the inspection standards" were assigned to a "high-pressure" group, while those who disagreed or strongly disagreed to this statement were assigned to the "low-pressure" group.

Table 2. Pressure to do well on the inspection standards by school types.

	I feel pressure to do well on the inspection standards									
	1	2	3	4	5	*N*	Mean	*SD*	*t* value $_{(df)}$	*p*
Primary	5.8%	9.5%	16.9%	41.0%	26.8%	639	3.73	1.127	−1.132 $_{(df = 1084)}$	.258
Secondary	7.1%	10.6%	20.6%	40.7%	21.0%	447	3.58	1.142		
Total	6.4%	10.2%	18.6%	40.8%	23.9%	1086	3.66	1.137		

Notes: Reply format: 1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*. Swiss data do not include information about the school type; thus, *t* test analysis is calculated without Swiss data.

For analyzing the influence of "accountability pressure" in national inspection systems (i.e., Research questions 3 and 4), only data sets from England, Sweden, and Austria were used, which will be explained in the respective chapter in more detail.

The development of the measurement model and all latent constructs are methodologically described in Gustafsson et al. (2013). For our analyses, nine latent variables were defined on the basis of 23 items (see Table A2 in Appendix 1 for an overview of all latent constructs); among those, a nested-factor model (Gustafsson & Balke, 1993) was constructed: "Improvement in capacity building" includes three narrow variables ("Improvement in teacher participation in decision making", "Improvement in teacher co-operation", "Improvement in transformational leadership"). In addition to descriptive and bivariate analyses, structural equation modeling (SEM) including multiple-group analyses (e.g., Byrne, 2010; Reinecke, 2005) were used to analyze the data. Model fit was estimated with standard measures: the chi-square goodness-of-fit test, the comparative fit index (CFI), and the root mean square error of approximation (RMSEA; e.g., Weiber & Mühlhaus, 2010). SPSS 20 and Amos 6.0 were used.

Results

Rating "accountability pressure"

In our discussion of inspection systems, "accountability pressure" was identified as a possible driving factor for producing the desired effects of inspections. Table 2 shows that a vast majority of nearly two thirds of school leaders in the overall sample report on feelings of "pressure to do well on inspection standards", while less than a fifth disagrees or strongly disagrees with this statement. No significant differences between primary and secondary school leaders were found.

Processes and effects of school inspection and "accountability pressure"

The first research question to be discussed is: Does "accountability pressure" promote development activities as a consequence of school inspections? If "accountability pressure" is a driving force for inspection impact, we might expect that "pressure" is an influence for improvement processes. Table 3 shows that this holds true for the improvement processes captured by our survey and supports our Hypothesis 1: Principals who feel more "accountability pressure" also say that more actions to improve self-evaluation, to

Table 3. Improvement processes by school leaders' experience of "accountability pressure".

	Pressure	*N*	Mean	*SD*	*t* value $_{(df)}$	*p*	*d*
Promoting/improving self-evaluation	Low	191	3.29	.705	−5.692 $_{(df = 321)}$	.000	.44
	High	696	3.62	.760			
Improvement in capacity building	Low	192	3.57	.561	−4.370 $_{(df = 280)}$	.000	.37
	High	702	3.76	.500			
Improvement in teacher participation in decision making	Low	173	3.61	.748	−2.807 $_{(df = 845)}$	.005	.24
	High	674	3.78	.711			
Improvement in teacher co-operation	Low	192	3.72	.707	−4.758 $_{(df = 289)}$	.000	.40
	High	700	3.99	.660			
Improvement in transformational leadership	Low	191	3.36	.603	−3.704 $_{(df = 889)}$	.000	.30
	High	700	3.55	.649			

Table 4. Intermediate mechanisms by school leaders' experience of "accountability pressure".

Intermediate mechanisms	Pressure	*N*	Mean	*SD*	*t* value $_{(df)}$	*p*	*d*
Setting expectations	Low	192	3.12	.826	−9.794 $_{(df = 251)}$	.000	.92
	High	749	3.74	.628			
Stakeholders sensitive to reports	Low	186	3.56	.885	−1.791 $_{(df = 249)}$	.074	.16
	High	741	3.68	.718			
Accepting feedback	Low	193	3.91	.826	1.289 $_{(df = 257)}$	.199	.12
	High	750	3.83	.656			

build development capacity among staff, to improve teacher participation in decision making, to enhance teacher co-operation, and to improve transformational leadership have been taken.

In a second step, we use the conceptual model of school inspection by Ehren et al. (2013) to explore intermediate processes which are meant to drive school improvement and various improvement actions. The model distinguishes three *intermediate inspection mechanisms*: expectations of good education set by the inspection standards, the behavior of key stakeholders, and schools accepting the inspection feedback. Table 4 indicates that school leaders who feel "high accountability pressure" also say more often that inspections are communicating expectations for quality criteria and work processes to them. There is, however, no difference with respect to the second intermediate mechanism: Principals who feel more or less "accountability pressure" do not see differences in their stakeholders' sensitivity to inspections. The third intermediate mechanism, "Accepting feedback", describes the assumption that inspections will promote intended effects through the feedback that is provided to schools during inspection visits and in inspection reports and through the fact that this feedback is understood, accepted, and eventually used by actors in schools to improve their work (see Ehren et al., 2013). Interestingly, headpersons who feel more "accountability pressure" are not more attentive and reactive to feedback than those who feel little. While this seems to run counter to the expectations which are captured in the "conceptual model", it is, however, consistent with the findings of Gustafsson et al. (2013; see, also, for Austria, Altrichter, Kemethofer, & Schmidinger, 2013). The explanation remains open: Possible measurement weaknesses of "Accepting feedback" and "Stakeholders' sensitivity" could be avoided by experimenting with alternative measures. An alternative explanation suggests that "feedback" could be a less effective driver of school improvement than conceived by the various versions of "evidence-based governance" (see Gustafsson et al., 2013).

In a third step, we turn to possible unintended effects of inspection: Table 5 shows that school leaders who experience more pressure also report significantly more unintended consequences. They react to "accountability pressure" by "discouraging new teaching

Table 5. Unintended consequences by school leaders' experience of "accountability pressure".

	Pressure	*N*	Mean	*SD*	*t* value $_{(df)}$	*p*	*d*
Discouraging new teaching methods	Low	192	1.86	.919	−4.805 $_{(df = 466)}$	.000	.45
	High	276	2.30	1.016			
Narrowing the curriculum and instructional strategies	Low	191	1.67	.889	−12.133 $_{(df = 465)}$	.000	1.07
	High	276	2.90	1.305			

methods" and by "narrowing the curriculum and instructional strategies" which represent undesired side effects of school inspections and accountability systems in general. Cohen's *d* values in Table 5 indicate a considerable effect size of "accountability pressure", in particular on activities "narrowing the curriculum and instructional strategies". Hence, Hypothesis 2 is supported.

Accountability pressure in various inspection systems

We turn to the second question: Are different national inspection systems characterized by differential amounts of "accountability pressure"?

Table 6 reveals that there are clear differences between countries with respect to the degree of "accountability pressure" experienced by headpersons. In most countries, more than half of the respondents agree or strongly agree that they feel pressure to do well on inspection standards, while much fewer principals (between 2% and 25%) disagree. Austrian and Swiss school leaders, however, seem to experience least inspection pressure: 36% of Austrian and 31% Swiss principals deny that there is such pressure,

In accordance with our hypothetical analysis of inspection features (see Table 1), the rank order is led by England followed by The Netherlands[3], which have the highest mean scores. Inspection systems which include more challenging elements, such as differentiated inspections, thresholds for distinguishing failing schools, sanctions for low-performing schools, and reports on individual schools to the general public, are met with more feelings of "accountability pressure" by principals. Interestingly, in these countries there is also high consensus among school leaders in their evaluation of pressure as indicated by low standard deviations.

Swedish principals experience nearly as much feelings of accountability pressure as their Dutch colleagues; however, there are slightly more persons disagreeing or/and undecided. This deviates from our expectations in Table 1, but may be in line with the claim that Sweden, after the reforms of the last decades, has moved away from traditional "Scandinavian governance arrangements" into the direction of high-stake accountability systems (Scheinin, 2013). While Irish and Czech school leaders occupy middle grounds, Austrian and Swiss headpersons say – also in line with our expectations in Table 1 – that they experience least "accountability pressure". Generally, in countries with a lower

Table 6. Pressure to do well on the inspection standards by countries.

	I feel pressure to do well on the inspection standards						
Country	1	2	3	4	5	Mean	*SD*
ENG	1.7%	0.0%	3.0%	32.8%	62.8%	4.54	.717
NL	2.2%	0.0%	8.9%	66.7%	22.2%	4.07	.720
SE	2.5%	4.2%	15.5%	54.6%	23.1%	3.92	.882
IE	4.1%	20.7%	18.2%	38.8%	18.2%	3.46	1.133
CZ	8.9%	6.7%	35.6%	44.4%	4.4%	3.29	.991
CH	14.5%	16.1%	25.8%	38.7%	4.8%	3.03	1.159
AT	14.1%	22.1%	32.2%	27.4%	4.0%	2.85	1.097
Total	*6.4%*	*10.2%*	*18.6%*	*40.8%*	*23.9%*	*3.66*	*1.137*

Notes: $N = 1169$; n(ENG) = 235; n(NL) = 45; n(SE) = 355; n(IE) = 121; n(CH) = 62; n(CZ) = 45; n(AT) = 298; $chi^2 = 479.468$, $df = 24$, $p = .000$. 1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*.

Table 7. Means of accountability pressure by country.

	Pressure				
Country	High		← →		Low
ENG	4.54				
NL		4.07			
SE		3.92			
IE			3.46		
CZ			3.29	3.29	
CH				3.03	3.03
AT					2.85

Notes: The table represents means of accountability pressure reported by school leaders in the respective countries. Figures within a column do no differ significantly; between columns there is a significant difference ($p \leq .05$; Duncan test).

average level of "accountability pressure", there is more variation in the principals' assessment of pressure, and standard deviations are higher.

Table 6 suggests that the feeling of "accountability pressure" cannot be exclusively understood as a personal feature of individual respondents or a characteristic of the specific school they work in. Rather, the national contexts and inspection systems seem to contribute to characteristic degrees of "accountability pressure". A multilevel analysis which is not reported here in detail attributed 37% of the explained variance of "accountability pressure" to the country level. An analysis of variance reproduced in Table 7 shows that there are significant differences between the amount of accountability pressure felt by school leaders, which support Hypothesis 3: School leaders in "high-stake inspection systems" (such as in England) say that they feel more "accountability pressure" than those in "low-stake systems" (such as in Austria and Switzerland). Only the "middle grounds" of our analysis in Table 1 do not fully conform to the expectations: The feeling of "accountability pressure" reported by Swedish school leaders is much closer to the figures in "high-stake inspection systems" than was expected; "accountability pressure" of Czech principals is closer to the Swiss "low-stake" conditions than was expected.

For further analyses, we included countries which conformed to the following criteria: Countries were (a) to represent different "accountability traditions" as indicated by our analysis in Table 1, (b) to be characterized by an empirically distinctive amount of "accountability pressure" (indicated by significant differences in Table 7), and (c) to include an adequate number of cases for further analyses. England, Sweden, and Austria fit these criteria. England represents a system with high "accountability pressure", in Sweden principals feel a little less, but still a considerable amount of "accountability pressure", while Austrian school leaders seem to live in a "low accountability pressure" system. Data from The Netherlands, Ireland, the Czech Republic, and Switzerland were excluded.

Inspection effects and processes under different contextual conditions

Do national inspection systems which are characterized by more "accountability pressure" promote more, less, or qualitatively different development activities and do they lead to more unintended consequences? This is the third question to be discussed.

Principals in the Austrian low-pressure system say that they have less frequently been active in all kind of development activities than those in Sweden and England (see

Table 8. Improvement process by accountability system.

Improvement process	Accountability system	*N*	Mean	*SD*	p ($\leq .05$)
Promoting/improving self-evaluation	Austria	482	3.19	.657	ENG > SE > AT
	Sweden	998	3.51	.570	
	England	217	4.03	.791	
Improvement in capacity building	Austria	496	3.52	.533	ENG ≈ SE > AT
	Sweden	1007	3.82	.438	
	England	221	3.82	.535	
Improvement in teacher participation in decision making	Austria	487	3.52	.711	ENG > SE > AT
	Sweden	1005	3.79	.636	
	England	220	3.95	.705	
Improvement in teacher co-operation	Austria	490	3.64	.702	ENG > SE > AT
	Sweden	1005	4.07	.590	
	England	220	4.18	.638	
Improvement in transformational leadership	Austria	482	3.36	.612	SE ≈ ENG > AT
	Sweden	998	3.60	.633	
	England	220	3.56	.713	

Note: The last column presents the results of the comparison of group means (Duncan test) and is to be read in the following way: ">" … the group mean is significantly bigger ($p \leq .05$) than …; "≈" … the group mean does not significantly differ ($p \leq .05$) from ….

Table 8). In many dimensions, English principals report more developmental activities than those in Sweden; however, there are no differences with respect to improvement actions in "capacity building" and in "transformational leadership".

We also wanted to know whether or not national inspection systems which are characterized by more "accountability pressure" on individual schools promote qualitatively different development activities. Table 8 suggests a rather uniform pattern with respect to the preferred development activities: Principals who feel more pressure say that they do more, and those who feel less pressure say that they do fewer improvement activities, but they do not qualitatively different things.

When it comes to the *intermediate processes* predicted by the "conceptual model", English principals who work in a system with high accountability pressure are more attentive to the expectations set by inspections and to the views and actions of stakeholders than those in Sweden, while the latter are more attentive than those in the Austrian low-pressure system (see Table 9). The results with respect to the mechanism "Accepting feedback" are somewhat different: In Sweden and Austria, principals say more often that they attend and react to the feedback provided by inspections than those under English "high-pressure" conditions.

A general picture seems to emerge which is in line with Hypothesis 4: Accountability systems characterized by more pressure are associated with more development activities. But there is also evidence from prior studies that "accountability pressure" might be associated with unintended effects, such as narrowing of the curriculum, discouraging new teaching strategies, and cheating. Table 10 shows a significant connection between unintended consequences and accountability systems which supports Hypothesis 5. School leaders in the Austrian "low-pressure accountability" system say less often that they take measures to discourage new teaching strategies and to narrow the curriculum than those in the English and the Swedish system, which are characterized by more accountability pressure. The Swedish principals discourage more often new teaching

Table 9. Intermediate inspection mechanisms by accountability system.

Intermediate mechanisms	Accountability system	*N*	Mean	*SD*	*p* (≤ .05)
Setting expectations	Austria	310	3.34	.563	ENG > SE > AT
	Sweden	357	3.55	.689	
	England	237	3.96	.634	
Stakeholders sensitive to reports	Austria	288	3.32	.614	ENG > SE > AT
	Sweden	351	3.56	.626	
	England	238	4.03	.626	
Accepting feedback	Austria	305	3.88	.661	AT ≈ SE > ENG
	Sweden	359	3.86	.606	
	England	238	3.71	.747	

Note: The last column presents the results of the comparison of group means (Duncan test) and is to be read in the following way: ">" … the group mean is significantly bigger ($p \leq .05$) than …; "≈" … the group mean does not significantly differ ($p \leq .05$) from ….

Table 10. Unintended consequences by accountability system.

Improvement process	Accountability system	*N*	Mean	*SD*	*p* (≤ .05)
Discouraging new teaching methods	Austria	302	1.91	.793	SE > ENG > AT
	Sweden	354	2.62	.868	
	England	235	2.15	.926	
Narrowing the curriculum and instructional strategies	Austria	299	1.87	.757	ENG > SE > AT
	Sweden	353	2.29	.909	
	England	235	3.08	1.251	

Note: The last column presents the results of the comparison of group means (Duncan test) and is to be read in the following way: ">" … the group mean is significantly bigger ($p \leq .05$) than …; "≈" … the group mean does not significantly differ ($p \leq .05$) from ….

strategies than the English ones, while the latter more often react with narrowing the curriculum.

The effect of pressure in inspection systems

The previous passages have shown that school leaders who are experiencing different amounts of "accountability pressure" also engage in different amounts of school development activities. However, this analytic strategy does not allow to make causal assumptions on the mechanisms and effects of pressure in different accountability systems and to understand their internal dynamics. Hence, we used a structural equation modeling approach with multiple-grouping to discuss our fourth research question, how "accountability pressure" impacts on the school improvement and the underlying mechanisms of inspection under varying contextual conditions.

Gustafsson et al. (2013) have translated the "conceptual model" of school inspection (Ehren et al., 2013) into a path model and largely confirmed its inherent assumptions. We use their measurement model, however, with some simplifications due to our modified analytic intentions: We are not testing the full conceptual model but focus on possible

Table 11. Goodness-of-fit indices.

Model	Chi^2	*df*	CFI	RMSEA	90%CI
Unconstrained	921.1	432	.932	.025	.022–.027
Metric invariance	1055.3	458	.917	.026	.024–.029
Scalar invariance	1982.9	494	.793	.040	.038–.042
Partial invariance	952.0	446	.929	.025	.023–.027

effects of pressure on core processes of school development. The fit of the model was good with a CFI of .929, a RMSEA of .025, and a chi square/*df* ratio of 2.13 (see Table 11). All tests of model fit fulfill common criteria (e.g., Van de Schoot, Lugtig, & Hox, 2012; Weiber & Mühlhaus, 2010). All values presented in Figure 2 refer to the measurement model. This model tests on invariance. Invariance should be achieved to allow straightforward interpretation; however, according to Van de Schoot et al. (2012; see, also, Byrne, 2010), valid inferences about differences between latent factors can be made with partial invariance as well. Full invariance is therefore only needed when comparing the sum scores. To achieve partial invariance, it is necessary that at least two loadings and intercepts are equally constrained across groups. In our model, partial invariance was achieved (ΔCFI < 0.01; Byrne, 2010), which allows comparing the relationship of the constructs between all groups in the structural equation model (Van de Schoot et al., 2012; Weiber & Mühlhaus, 2010).

Our multiple-group analysis (see Figure 2) indicates that "accountability pressure" influences the "expectations set by inspections" in all accountability systems of our study. The effect of "accountability pressure" on quality expectations is smallest in the Austrian "low-accountability pressure system"; in Sweden (which is characterized by medium to high accountability pressure) and in the high-pressure accountability system of England, the effect of pressure on expectations is significantly higher. There is no direct influence

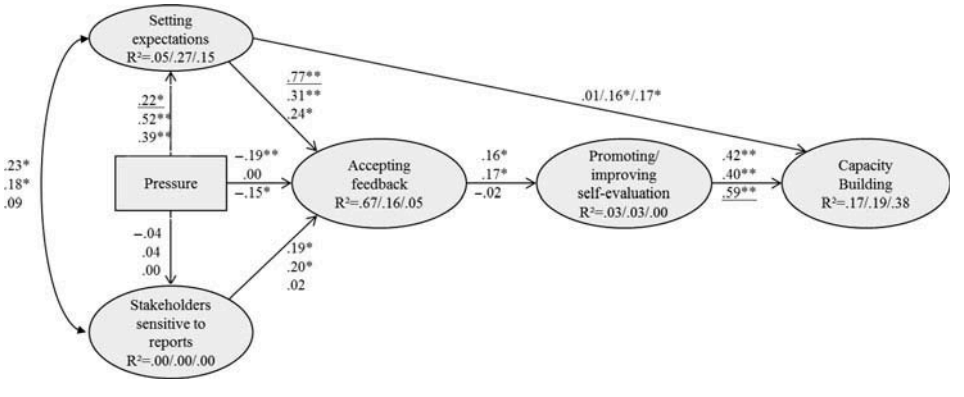

Figure 2. Effects of school inspection in accountability systems.
Notes: For every path, three estimates of path coefficients are included: The first one refers to Austria (low accountability pressure), the second one to Sweden (medium to high accountability pressure), and the third one to England (high accountability pressure). An underlined value indicates that this group differs significantly ($p \leq 05$) from other groups. The coefficients represent standardized direct effects.
*$p \leq .05$. **$p \leq .001$; n(AT) = 298, n(SE) = 355, n(ENG) = 235.

of "accountability pressure" on the attention headpersons pay to stakeholders and their reactions to inspections. However, "expectations set by inspections" and "sensitivity to stakeholders" are significantly correlated in Austria and Sweden, but not under English "high-pressure" conditions. In Austria and England, "accountability pressure" has a negative effect on the "acceptance of feedback", while in Sweden there is no effect. However, the "expectations set by inspections" significantly influence the school leaders' preparedness to "accept feedback" in every country.

What are the effects of these intermediate mechanisms on *improvement processes*? With the exception of England, "improving self-evaluation" is significantly influenced by the "acceptance of inspection feedback". If schools are "improving self-evaluation", this will trigger other school development activities under all accountability conditions, but the effect is significantly higher in the English high-pressure system. Attention to "stakeholder's reactions" significantly promotes the "acceptance of inspection feedback" in Austria and Sweden, but not so in England. The "expectations set by inspection" seem to directly influence activities to improve a school's development capacity only in Sweden and England. In the Austrian "low-stakes system", there is no significant path from "setting expectations" to "capacity building".

These results cast some doubt on the assumptions of inspection impact captured in the "conceptual model", which are subscribed to by authorities in all accountability systems of our study. Compared to the results of Gustafsson et al. (2013), it is evident that the addition of "accountability pressure" to the measurement model results in a more differentiated account, since the hypothesized effective paths cannot be established for all accountability systems in the same way. Eventually, it turns out that the simplification of the measurement model does not massively damage the explanative power of the model; the proportion of explained variance is between 17% and 38% for different accountability systems.

Summary and discussion

The empirical parts of the paper explored the role of "accountability pressure", which school leaders attribute to school inspections for various activities of school improvement. First, we have found a connection between "accountability pressure" and the amount of development activities as was expected in Hypothesis 1: School leaders who feel more "accountability pressure" say that more development activities take place in their schools. The situation is less clear with respect to the intermediate processes found by Ehren et al. (2013): Headpersons who feel more pressure to do well on inspection standards are more sensitive to the expectations communicated by inspection systems; however, they do not differ in their attention to stakeholders' reactions to inspection. And there is no differential reaction to inspection feedback, which sheds some doubt on the hypothesized role of "feedback" in governance processes. Our results also show that school leaders who experience more "accountability pressure" see significantly more unintended consequences, which supports our second hypothesis.

In a second step, these findings are specified for different inspection systems. Our data show that school leaders feel that different degrees of "accountability pressure" are included in different national inspection systems. As had been predicted by an analysis of inspection characteristics and by Hypothesis 3, most pressure is reported by English and least pressure by Swiss and Austrian principals. Swedish headpersons' experiences of "accountability pressure" resemble more closely those in high-stake inspections systems than was expected.

Third, bivariate analyses indicate – in line with Hypotheses 4 and 5 – that school leaders in a system characterized by "high accountability pressure" such as the English are consistently more active with respect to improvement processes and are more attentive to inspection expectations and to stakeholders' reactions. Swedish principals who feel less accountability pressure than the English headpersons but still much more than Austrian principals are in some respects between these two groups, in other respects their reactions resemble those of the English "high-pressure" context.

The analysis so far seems to suggest a clear-cut formula: The more "accountability pressure", the more development one gets. However, quantitative increase of development activities need not necessarily include quality gains. In line with Hypothesis 5, we found an indication that in systems with "high inspection pressure" also the amount of unintended consequences is increasing: School leaders in the Austrian "low-pressure accountability" system say less often that they take measures to discourage new teaching strategies and to narrow the curriculum than those in the English and the Swedish systems, which are characterized by more accountability pressure.

Fourth, a SEM multigroup analysis was performed to clarify how pressure impacts on school improvement and the underlying mechanisms of inspection under varying contextual conditions. We found a consistent effect of "accountability pressure" on the "quality expectations" communicated through inspection systems, but it seems to vary with the amount of pressure of the specific governance regimes. It is smallest under Austrian low-pressure conditions and significantly higher in Sweden and England. "Quality expectations" and the "sensitivity to stakeholders' reactions" are significantly correlated in Austria and Sweden; however, there is no direct influence of "accountability pressure" on the "attention principals pay to stakeholders" in any accountability system. The hypothesized effective paths of inspection impact (which have been claimed by inspection representatives; see Ehren et al., 2013) cannot be established for all accountability systems in the same way. "Improving self-evaluation" is significantly influenced by "acceptance of inspection feedback" in Austria and Sweden, but not so in the English "high-pressure" context. "Improving a school's development capacity" is directly influenced by "quality expectations" only in Sweden and England, but not so in Austria. The most consistent path seems plausible but trivial: Those schools which engage in "improving self-evaluation" will also engage in "improving capacity building".

What conclusions can be drawn from this account? First, "accountability pressure" generally seems to be an important driving factor in present inspection systems: It influences the major intermediate mechanism "Setting expectations" and is associated with most of those development activities which inspection systems intend to stimulate. Second, interpretation must take into account that we work with self-reported data of school leaders. When school leaders report more developmental activities in situations with high accountability pressure, this may be due to the fact that they feel more "pressure" to provide such reports. Whether or not these reports actually reflect more developmental activities must be the object of further investigations. Data about school quality, such as external assessment of individual schools' quality and information on student performance, would allow substantiating the argument.

Third, inspection systems seem to differ in the amount of "accountability pressure" they put on principals. While we have seen that the degree of "accountability pressure" is positively associated with more development activities under all contextual conditions in our study, we have also some evidence that there may be, at least in part, differential processes underlying the relationship of inspection and school improvement (see the discussion of Figure 2 in the section above), which calls for more detailed analytic

approaches. Fourth, we have used a broad-brush characterization of inspection systems by sum scores of "accountability pressure felt by principals". Further studies should have a closer look at what elements and constellations of inspections stimulate feelings of "accountability pressure" and subsequent development activities.

Fifth, "pressure" may come from various sources (e.g., from state arrangements such as inspection systems, from competitive constellations, from parent wishes, from high internal aspirations). It is obviously one reason which stimulates development. However, one would not recommend both a comprehensive theory of educational effectiveness and a governance system of education to rely fully on this mechanism, since "accountability pressure" is also connected with undesirable consequences (such as narrowing the curriculum and discouraging teaching innovation) which may contradict and inhibit the inspections' quality aims.

Present inspection systems and, in particular, "high-stake inspection systems" seem to work on pressure, but not exclusively so – as we can explain only between 17% and 38% of the variance of the developmental target activities of "capacity building". For the advancement of the educational effectiveness theory, further studies must check possible alternative and/or complementary driving mechanisms for their relative weight and for potential collusion or interference with the forces of "accountability pressure". A possible lead to alternatives is given by Böttger-Beer and Koch (2008; see, also, Bryk & Schneider, 2002, for another alternative). They distinguish three different modes of development by evaluative measures which are associated with distinct driving forces: Development may be driven by *competition* when evaluation information is made publicly available and used by parents and students for school choice and by schools for devising competitive improvement measures. Development may be driven by *external intervention* when evaluation information is mainly used by the authorities to stimulate improvement through sanctions or through special support measures. Development may also be driven by *insight* when evaluation information is mainly used by the internal professionals of individual schools to better understand the specific situation of a school and to devise worthwhile options for improvement. The authors claim that their system of "external evaluation" in the German state of Saxonia is effective through the mechanisms of insight and supportive intervention and that it abstains from *development by pressure* based on the mechanisms of competition and sanctions. While systems which emphasize "insight" seem to conceive development as a primarily informational problem, those investing in "pressure" see it as a motivational one. Thus, further studies should investigate the role of "insight" and other "driving mechanisms" in existing inspection arrangements; they should also include inspection systems which claim to be based mainly on such alternative mechanisms into their data base.

Notes

1. Data collection was carried out by an international team of researchers, including (in addition to the authors of this paper) Melanie Ehren, Gerry Conyngham, David Greger, Jan-Eric Gustafsson, Stephan Gerhard Huber, Karen Jones, Gerry McNamara, Eva Myrberg, Joe O'Hara, Guri Skedsmo, and Peter Tymms.
2. The Swiss part of the project is funded by Stiftung Mercator Schweiz.
3. Interpretation of Dutch data must take into account the caveat discussed in the section on data sources and methods.

References

Altrichter, H., & Heinrich, M. (2007). Kategorien der Governance-Analyse und Transformationen der Systemsteuerung in Österreich [Categories of governance analysis and transformations of

system governance in Austria]. In H. Altrichter, T. Brüsemeister, & J. Wissinger (Eds.), *Educational Governance* (pp. 55–103). Wiesbaden: VS Verlag für Sozialwissenschaften.
Altrichter, H., & Kemethofer, D. (2012, September). *School leaders and their experience of school inspection*. Paper presented at the European Conference on Educational Research, Cadiz.
Altricher, H., Kemethofer, D., & Schmidinger, E. (2013). Neue Schulinspektion und Qualitätsmanagement im Schulsystem [New school inspection and quality management in schooling]. *Erziehung und Unterricht*, *163*, 961–978.
Bellmann, J., & Weiß, M. (2009). Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem [Risks and side effects of new governance in the school system]. *Zeitschrift für Pädagogik*, *55*, 286–308.
Böttger-Beer, M., & Koch, E. (2008). Externe Schulevaluation in Sachsen – ein Dialog zwischen Wissenschaft und Praxis [External school evaluation in Saxony – a dialogue between science and practice]. In W. Böttcher, W. Bos, H. Döbert, & H. G. Holtappels (Eds.), *Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive* (pp. 253–264). Münster: Waxmann.
Braithwaite, J. (2008). *Regularity capitalism. How it works, ideas for making it work better*. Cheltenham: Edward Elgar.
Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource of improvement*. New York, NY: Russel Sage Foundation.
Byrne, B. M. (2010). *Structural equation modeling with AMOS. Basic concepts, applications, and programming*. New York, NY: Routledge.
Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, *93*, 1045–1057.
Creemers, B. P. M., & Kyriakides, L. (2012). *Improving quality in education. Dynamic approaches to school improvement*. London: Routledge.
Creemers, B. P. M., Stoll, L., & Reezigt, G. (2007). Effective school improvement – ingredients for success: The results of an international comparative study of best practice case studies. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement* (pp. 825–838). New York, NY: Springer.
Cullingford, C., & Daniels, S. (1999). Effects of OFSTED inspections on school performance. In C. Cullingford (Ed.), *An inspector calls* (pp. 59–69). London: Kogan Page.
Dedering, K., & Müller, S. (2011). School improvement through inspections? First empirical insights from Germany. *Journal of Educational Change*, *12*, 301–322.
De Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education*, *33*, 379–396.
Ehren, M. C. M., Altrichter, H., McNamara, G., & O'Hara, J. (2013). Impact of school inspections on improvement of schools – describing assumptions on causal mechanisms in six European countries. *Educational Assessment, Evaluation and Accountability*, *25*, 3–43.
Ehren, M. C. M., & Swanborn, M. S. L. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement*, *23*, 257–280.
Ehren, M. C. M., & Visscher, A. J. (2008). The relationships between school inspections, school characteristics and school improvement. *British Journal of Educational Studies*, *56*, 205–227.
Elstad, E., Christophersen, K.-A., & Turmo, A. (2012). The strength of accountability and teachers' organisational citizenship behaviour. *Journal of Educational Administration*, *50*, 612–628.
Eurydice. (2004). *Evaluation of schools providing compulsory education in Europe*. Retrieved from http://www.eurydice.org/portal/page/portal/Eurydice
Faubert, V. (2009). *School evaluation: Current practices in OECD countries and a literature review* (OECD Educational Working Papers, No. 42). Retrieved from http://dx.doi.org/10.1787/218816547156
Good, T. L., Wiley, C. R. H., & Sabers, D. (2010). Accountability and educational reform: A critical analysis of four perspectives and considerations for enhancing reform efforts. *Educational Psychologist*, *45*, 138–148.
Greger, D. (2011). *School inspection of Czech schools: A critical reflection on intended effects and causal mechanisms*. Unpublished working paper, LLP project "Impact of school inspections on teaching and learning", Charles University, Prague.
Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407–434.

Gustafsson, J.-E., Ehren, M. C. M., Conyngham, G., McNamara, G., Altrichter, H., & O'Hara, J. (2013). *School inspections and school improvement: Testing assumptions on causal mechanisms*. Manuscript submitted for publication.

Gustafsson, J.-E., & Myrberg, E. (2011). *School inspections of Swedish schools: A critical reflection on intended effects, causal mechanisms and methods*. Unpublished working paper, LLP project "Impact of school inspections on teaching and learning", University of Gothenburg, Gothenburg.

Hanushek, E. A., & Raymond, M. E. (2002). *Lessons about the design of state accountability systems*. Retrieved from http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BRaymond%202003%20NCLB.pdf

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management 24*, 297–327.

Huber, S. G. (2011). School governance in Switzerland: Tensions between new roles and old traditions. *Educational Management Administration & Leadership*, *39*, 469–485.

Hughes, G., Mears, R., & Winch, C. (1997). An inspector calls? Regulation and accountability in three public services. *Policy & Politics*, *25*, 299–313.

Husfeldt, V. (2011). Wirkungen und Wirksamkeit der externen Schulevaluation. Überblick zum Stand der Forschung [The impact of school inspection – does it really work? State of research]. *Zeitschrift für Erziehungswissenschaft*, *14*, 259–282.

Klerks, M. (2013). *The effect of school inspections: A systematic review*. Manuscript submitted for publication.

Kotthoff, H.-G. (2003). *Bessere Schulen durch Evaluation?* [Better schools through evaluation?]. Münster: Waxmann.

Kotthoff, H.-G., & Böttcher, W. (2010). Neue Formen der "Schulinspektion": Wirkungshoffnungen und Wirksamkeit im Spiegel empirischer Bildungsforschung [New forms of "school inspection": Empirical education research on expectations of effects and effectiveness]. In H. Altrichter & K. Maag Merki (Eds.), *Handbuch Neue Steuerung im Schulsystem* (pp. 295–325). Wiesbaden: VS Verlag für Sozialwissenschaften.

Kotthoff, H.-G., Maag Merki, K., & Böttcher, W. (2007). Schulinspektion im internationalen Vergleich: Konzepte und Erfahrungen – England, Niederlande, Schweden, Schweiz [International comparison of school inspection: Concepts and experiences – England, Netherlands, Sweden, Switzerland]. *Journal für Schulentwicklung*, *3*, 53–58.

Lee, J. (2006). Input-guarantee versus performance-guarantee approaches to school accountability: Cross-state comparisons of policies, resources, and outcomes. *Peabody Journal of Education*, *81*(4), 43–64.

Lindgren, J., Hult, A., Segerholm, C., & Rönnberg, L. (2012). Mediating school inspection – key dimensions and keywords in agency text production 2003–2010. *Education Inquiry*, *3*, 569–590.

Luginbuhl, R., Webbink, D., & De Wolf, I. (2009). Do school inspections improve primary school performance? *Educational Evaluation and Policy Analysis*, *31*, 221–237.

Maag Merki, K. (2010). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen [Theoretical and empirical analyses of the efficiency of educational standards, standard-related tests and central final examinations]. In H. Altrichter & K. Maag Merki (Eds.), *Handbuch Neue Steuerung im Schulsystem* (pp. 145–169). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maag Merki, K. (2011). Special issue: Accountability systems and their effects on school processes and student learning. *Studies in Educational Evaluation*, *37*, 177–179.

Maritzen, N. (2007). Schulinspektion – ein neues Element der Systemsteuerung [School inspection – a new element of system governance]. *Journal für Schulentwicklung*, *3*, 8–14.

Matthews, P., & Sammons, P. (2004). *Improvement through inspection. An evaluation of the impact of Ofsted's work*. London: Ofsted.

Matthews, P., & Sammons, P. (2005). Survival of the weakest: The differential improvement of schools causing concern in England. *London Review of Education*, *3*, 159–176.

McCrone, T., Rudd, P., Blenkinsop, S., Wade, P., Rutt, S., & Yeshanew, T. (2007). *Impact of Section 5 inspections: Maintained schools in England: Final report*. Retrieved from http://www.ofsted.gov.uk/resources/impact-reports-2007

Murphy, J. (2009). Turning around failing schools: Policy insights from the corporate, government, and nonprofit sectors. *Educational Policy*, *23*, 796–830.

Ozga, J., & Jones, R. (2006). Travelling and embedded policy: The case of knowledge transfer. *Journal of Education Policy*, *21*, 1–17. doi:10.1080/02680930500391462
Perryman, J. (2007). Inspection and emotion. *Cambridge Journal of Education*, *37*, 173–190.
Perryman, J. (2010). Improvement after inspection. *Improving Schools*, *13*, 182–196.
Reed, E., Scull, J., Slicker, G., & Winkler, A. M. (2012). *Defining strong state accountability systems. How can better standards gain greater traction? A first look*. Retrieved from http://files.eric.ed.gov/fulltext/ED532505.pdf
Reezigt, G. J., & Creemers, B. P. M. (2005). A comprehensive framework for effective school improvement. *School Effectiveness and School Improvement*, *16*, 407–424.
Reinecke, J. (2005). *Strukturgleichungsmodelle in den Sozialwissenschaften* [Structural equation modeling in social science]. München: Oldenbourg.
Rosenthal, L. (2004). Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review*, *23*, 143–151.
Scheinin, P. (2013, March). *Does evidence influence policy?* Paper presented at the symposium "Ideas and consequences of market principles in education: The Swedish case", Royal Swedish Academy of Sciences, Stockholm.
Scheipl, J., & Seel, H. (1985). *Die Entwicklung des österreichischen Schulwesens von 1750 bis 1938* [The development of the Austrian education system – 1750 to 1938]. Graz: Leykam.
Shaw, I., Newton, D. P., Aitkin, M., & Darnell, R. (2003). Do Ofsted inspections of secondary schools make a difference to GCSE results? *British Educational Research Journal*, *29*, 63–75.
Skedsmo, G., & Huber, S. G. (2013). *Securing educational quality through different cantonal approaches of external evaluation*. Manuscript in preparation.
Valverde, G. A., & Schmidt, W. H. (2000). Greater expectations: Learning from other nations in the quest for "world-class standards" in US school mathematics and science. *Journal of Curriculum Studies*, *32*, 651–687.
Van Bruggen, J. C. (2010). *Inspectorates of education in Europe; some comparative remarks about their tasks and work* (SICI report). Retrieved from http://www.sici-inspectorates.eu/getattachment/c2bfe3ff-49b7-4397-ae65-d0a203451928
Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Development Psychology*, *9*, 486–492.
Weiber, R., & Mühlhaus, D. (2010). *Strukturgleichungsmodellierung* [Structural equation modeling]. Heidelberg: Springer.

Appendix

Table A1. Summary of sample characteristics in each country.

	Target Sample		Actual Sample		Response rate	
Country	Primary	Secondary	Primary	Secondary	Primary	Secondary
Netherlands (NL)	408	359	73	15	18%	4%
England (ENG)	1422	637	189	101	13%	16%
Sweden (SE)	1167	987	567	464	49%	47%
Ireland (IE)	3200	729	123	42	4%	6%
Czech Republic (CZ)	150	170	56	69	37%	41%
Austria (AT)	503	194	345	149	69%	77%
Switzerland (CH)[1]	465		132		28%	

Note: [1]In Switzerland, no information about school type in the target sample is available. Three principals responded that they work in a special school, and four did not respond to this question.

Table A2. Overview of latent variables.

Latent construct	Example item	Number of items	Scale	Cronbachs Alpha	*N*
Setting expectations	The inspection standards affect the evaluation and supervision of teachers.	7	*strongly agree* (1) – *strongly disagree* (5)	.84	1109
Stakeholders sensitive to reports	The school's Board of Management/Boards of Governors is very aware of the contents of the school inspection report.	3		.71	1112
Accepting feedback	The feedback received from the inspectors was useful.	5		.81	1218
Promoting/improving self-evaluation	Compared to last academic year, I spent less/more time on the self-evaluation process as a whole.	3	*much less* (1) – *much more* (5)	.87	2316
Improvement in capacity building	Compared to last academic year, I spent less/more time involving teachers in making decisions about using new teaching methods.	7		.80	2203
Improvement in teacher participation in decision making[1]	Compared to last academic year, I spent less/more time involving teachers in making decisions about using new teaching methods.	2		.63	2280
Improvement in teacher co-operation[1]	Compared to last academic year, I spent less/more time on improving teachers' collaboration in discussing assessment results of students with each other.	2		.68	2364
Improvement in transformational leadership[1]	Compared to last academic year, I spent less/more time in my leadership role communicating the school's vision to the staff, pupils, parents and others.	2		.63	2316

Note: [1]subset of items representing a narrow latent variable in the nested-factor "Improvement in capacity building" (see Gustafsson et al., 2013).