Studia paedagogica
vol. 22, n. 2, 2017
www.studiapaedagogica.cz
https://doi.org/10.5817/SP2017-2-2

DELPHI STUDY ON STANDARDIZED SYSTEMS TO MONITOR STUDENT LEARNING OUTCOMES IN FLANDERS: MECHANISMS FOR BUILDING TRUST AND/OR CONTROL?

MARTEEN PENNINCKX, AMY QUINTELIER,
JAN VANHOOF, SVEN DE MAEYER,
PETER VAN PETEGEM

Abstract

Several countries have implemented monitoring systems where students need to take standardized tests at regular intervals. These tests may serve either a development-oriented goal that supports public trust in schools, or a more accountability-oriented perspective to increase control. Currently, the Flemish education system has no standardized testing. The idea of implementing a monitoring system is highly contentious. By means of a Delphi study with policy makers, education specialists, school governors, principals, teachers, and a student representative (n=24), we identified the characteristics of a monitoring system that would be accepted by different stakeholders. Based on these characteristics, we proposed eight scenarios for future policy development. Next, the desirability of these scenarios was assessed by each respondent. The results show that in order to gain broad social support, a focus on strengthening trust is preferred over a focus on control through such measures as avoiding the public availability of test results. In addition, other key results for the development and implementation of a system to monitor student learning outcomes are discussed.

Keywords

standardized tests, Delphi study, learning outcomes, learning progress, added value, policy scenarios, monitoring system

Introduction

Improving the quality of education is a permanent concern for national policy makers. In that regard, a growing number of countries have implemented monitoring systems where students need to take standardized tests at regular intervals during their school career (OECD, 2013). These tests may inform teachers and schools about the strengths and weaknesses of their education in light of the curricula, supporting schools in maximizing the fit between school/classroom processes and student learning needs. As a result, public trust in schools may be endorsed. Systems for monitoring student learning outcomes may also play a role in the certification of students or in the external evaluation of schools or individual teachers. In this way, standardized monitoring systems are used for accountability and/or control perspective (Vanhoof & Van Petegem, 2007; Wang, Beckett, & Brown, 2006).

In many education systems, there is an ongoing debate about the desirability of different mechanisms to monitor student learning outcomes (OECD, 2013). There is ample evidence of positive effects as well as undesirable side effects from systems based on standardized tests (Au, 2007; Wiliam, 2010). The occurrence of these (side) effects is often linked to the monitoring system's aims (trust or control). This debate is embedded in a larger "clash of philosophical positions" with different views on educational quality and a "clash of interests" between different stakeholders (Phelps, 2005).

In contrast to its neighboring regions and countries, the Flemish education system does not have any standardized testing of students. The idea to implement such a system to monitor student learning outcomes is highly contentious (Shewbridge, Hulshof, Nusche, & Stoll, 2011). The discussion needs to be seen within the context of the constitutional principle of freedom of education, which grants great autonomy to Flemish schools. Traditionally, there has always been trust in the competency of schools to provide quality education. To an increasing extent, however, there are demands to hold schools accountable for their educational processes as well as to ensure the quality of schools' self-evaluations and teachers' assessments of students in order to ensure public trust in schools.

Both demands might be addressed by the implementation of a system to monitor student learning outcomes. This study starts from the following research questions:

1. According to Flemish stakeholders, what are characteristics of a desirable system to monitor student learning outcomes?
2. How can these characteristics be translated into scenarios for future policy development?
3. How are these scenarios evaluated by Flemish stakeholders with regard to desirability?

This study aims to identify the characteristics of a monitoring system that would be accepted by different stakeholders. The opinions of different stakeholders in the Flemish educational context with regard to the above questions had not previously been mapped in a systematic manner. Therefore, identifying and incorporating these opinions is a task of an exploratory nature.

Theoretical framework

Although there is ample literature on the effects of standardized systems, it remains difficult to identify clear guidelines for the design of an effective monitoring system. This is mainly due to two obstacles: (1) the fact that a great deal of the literature is centered on the effects of high-stakes exit exams organized from an accountability-oriented control perspective, while there is less evidence about the effectiveness of other sorts of monitoring systems that aim to contribute to public trust in schools; and (2) various author's lack of distinction between opinions and empirical results.

Effectiveness of standardized systems in monitoring student learning outcomes

Studies have shown that standardized tests give teachers a more objective view about the learning outcomes of their students. Teacher assessments are often (unconsciously) influenced by irrelevant factors and beliefs (Baird, Ahmed, Hopfenbeck, Brown, & Elliott, 2013; Black, Harrison, Hodgen, Marshall, & Serret, 2011; Marlow et al., 2014). Consequently, standardized tests hold greater promise for students in different schools to be assessed in a similar way and assessed on an equivalent standard in order to be awarded a certificate or diploma (Neumann, Trautwein, & Nagy, 2011). Standardized tests may therefore serve as a means to increase control over the quality of school-based assessments, but may also strengthen public trust in schools. Several scholars have found a gap between scores awarded by teachers (through their own tests) and scores awarded by standardized tests (De Lange & Dronkers, 2007; Schildkamp, Rekers-Mombarg, & Harms, 2012). In contrast with the general interpretation that standardized tests provide a more objective picture of students' knowledge and skills, opponents have argued that standardized tests are limited in their assessment capacity as they can take only a small number of elements into account, while teachers have a "richer" view of the capacity of their students (Allal, 2013; Haertel, 1999). In this way, standardized tests may provide incomplete and even unfair judgment about students' capacities, which may explain the discordance with scores awarded by teachers (Gipps & Stobart, 2009; Klenowski, 2014). Moreover, through their format standardized tests may also be negatively

biased against certain groups of students (e.g., non-native language speakers) (Brennan, Kim, Wenz-Gross, & Siperstein, 2001; Carnoy & Loeb, 2002; Cobb & Russell, 2014).

Another area of disagreement relates to the (assumed) effect that standardized tests may lead to improved learning outcomes (Bishop, 1998; Klein & Van Ackeren, 2012). Although several studies have provided evidence for this hypothesis (Jürges, Büchel, & Schneider, 2005; Jürges, Richter, & Schneider, 2005; Wössmann, 2005), other scholars opine that the gain is merely the result of teaching to the test, with teachers focusing on the curricular areas and lower-order skills that are typically tested by standardized tests at the expense of other curricular areas or higher-order skills which are more difficult to evaluate (Au, 2007; Klein, Hamilton, McCaffrey, & Stecher, 2000; Wang et al., 2006). Research has also shown that when monitoring systems lead to improved learning outcomes, they may result in a negative impact on autonomous student motivation (Jürges & Schneider, 2010; Van Ackeren, Block, Klein, & Kühn, 2012).

Other assumed effects from standardized monitoring systems include a positive effect on teaching methods (Van Ackeren et al., 2012) and a wash back effect on teachers' own tests (Cizek, 2005). Negative side effects from a monitoring system may include the strong emotional impact that standardized tests (particularly exit exams) have on students (Jürges, Schneider, Senkbeil, & Carstensen, 2009; Segool, Carlson, Goforth, von der Embse, & Barterian, 2013), a potential negative impact on students' study careers (Amrein-Beardsley, Berliner, & Rideau, 2010; Haney, 2000), undesirable strategic behavior from students (Van Ackeren et al., 2012), and a negative impact on teachers' feeling of professionalism and motivation (Shepard, 1992; Smagorinsky, Lakly, & Johnson, 2002; Van Ackeren et al., 2012).

It is notable that the empirical evidence is limited to students and teachers. While an (assumed) impact on school policy (e.g., schools' self-evaluations) has often been claimed (Keeves, Hungi, & Afrassa, 2005; Loeb, 2013; Saunders, 1999), this impact is not (yet) corroborated by empirical studies. In contrast, there is evidence about standardized tests' negative effect at the school level, particularly when student learning outcomes are used in school accountability and when rankings are published based on test results (Collins, Reiss, & Stobart, 2010; Karsten, Visscher, & De Jong, 2001; Perryman, Ball, Maguire, & Braun, 2011).

Characteristics of effective standardized systems

There are large differences in the conceptualization of different standardized monitoring systems. Several scholars believe that the effectiveness of such systems depends largely on the question of whether they enable conclusions to be drawn about student progress (Janssens, Rekers-Mombarg, & Lacor,

2014; Popham, 2005) and added value for schools (Loeb, 2013). In addition, the tests' reliability and validity (Sireci, 2005; Wiliam, 2010), close or loose links with the curriculum (Goodman & Hambleton, 2005; Shepard, 1992; Wang et al., 2006), and the extent of teacher involvement in test development are mentioned as influencing the systems' effectiveness. Contested issues include the height of the stakes in the tests, schools' autonomy to decide on participation, and testing frequency.

The effectiveness of standardized monitoring systems does not depend only on the system itself, but also on qualities of schools (e.g., the schools' "data literacy" [Beaver & Weinbaum, 2015] or the tests' acceptance within the school [Ramsteck et al., 2015; Saunders, 2000]) and characteristics of the educational system in general, such as its preparedness to invest into professional development related to the use of the monitoring system's results (Tymms, 1997; Wiliam, 2010).

Method

This study aims to identify the characteristics of a monitoring system that can count on a broad base of social support within the Flemish education system. Such characteristics cannot simply be deduced from the international empirical literature. Therefore, this study began by identifying opinions and points of view held by different stakeholders in the Flemish education context.

The exploratory nature of this aim supports the choice of a Delphi study. Through the technique of iterative feedback, this method enables exploration of and confrontation among the opinions of various experts (Dalkey & Helmer, 1963). It is essentially a technique to structure the process of group communication in such a way as to allow different experts to reflect on a complex matter (Day & Bobeva, 2005; Linstone & Turoff, 1975). Delphi studies generally serve a dual goal: to provide a reliable and creative exploration of different points of view on the one hand, and to gather adequate information on clearly aligned policy issues on the other (Adler & Ziglio, 1996).

One of the key issues in setting up a Delphi study is defining proper criteria for the selection of respondents (Chong, Adnan, & Zin, 2012; Okoli & Pawlowski, 2004). Each respondent needs to have substantial knowledge about the topic and be able to create and clarify an opinion autonomously (Day & Bobeva, 2005). In line with recommendations from the literature (Hsu & Sandford, 2007), policy stakeholders as well as users (teachers, school principals) and evaluees (students) were included in the expert panel. Four categories of respondents were selected to create a panel of 24 experts (see Table 1). The category *policy makers* includes civil servants at the Flemish

Ministry of Education and Training and the Inspectorate of Education; *education providers* includes those providers' school counselors; *schools* comprises principals and teachers at primary and secondary schools and a staff member of a student representative body; and *experts* includes academics, opinion makers, and respondents with expertise in the development of standardized tests.

Table 1
Expert panel in four categories

Category	No. of respondents
Policy makers	4
Education providers	6
Schools	8
Experts	6
Total	24

The respondents remained anonymous throughout the process, and so respondents did not know one another's identities (Dalkey & Helmer, 1963). This Delphi study consisted of three research rounds. In the first round, each of the 24 respondents was given a questionnaire with open questions. The questionnaire had a very broad nature, including such questions as "In your opinion, what are the advantages for students from standardized monitoring of student learning outcomes and/or learning progress?"; "In your opinion, what are the disadvantageous side effects from standardized monitoring of student learning outcomes and/or learning progress?"; and "In your opinion, what are the requirements to strengthen the advantages and avoid the disadvantages?". The data resulting from this written questionnaire were analyzed deductively without preexisting categories in line with the principles of grounded theory (Strauss & Corbin, 1990). The first round resulted in *a theory grounded in data* about the assumed effects and side effects of standardized systems to monitor student learning outcomes and the conditions under which these effects are assumed to occur.

During the second round, the point of view of each respondent was elaborated and refined by means of an in-depth interview. Each respondent was also given the chance to respond to other respondents' ideas formulated in the first round. Questions included "Imagine Flanders implemented standardized tests next year, with or without student progress monitoring; in this case, what should be the goal of this monitoring of student learning outcomes?"; and "When students have taken standardized tests, what information should be available for a) schools, b) students, and c) authorities?".

This data collection enabled elaboration of the grounded theory from the first round, and key elements as well as areas of dissension were identified. Data were coded following the principles of the framework approach (Ritchie & Spencer, 1993).

Based on the results of the first and second rounds, the research team combined several key elements into eight different policy scenarios. Scenarios are a helpful tool to facilitate communication and foster reflection by respondents on issues or points of view that had not been previously discussed (Mietzner & Reger, 2005). These scenarios were written to achieve the maximum variety of different conceptualizations of standardized systems to monitor student learning outcomes. During the process of scenario writing, we adhered to the principles of scenario development as described by Schwartz (1991).

Each scenario consists of a description of a theoretical system to monitor student learning outcomes, including its expected implications. In the third round, all eight policy scenarios were sent to each respondent. The respondents scored each scenario on its desirability (a Likert scale on the item "I think this scenario is desirable" ranging from 1 = "I totally disagree" to 5 = "I totally agree"). Then, each respondent was interviewed a second time in order to gain additional information about the arguments underlying their judgments. The analysis framework was built on data from the first and second research rounds as well as the eight policy scenarios. None of these frameworks were known at the start of the study.

Appendix 2 provides some example quotes resulting from the research rounds.

Results

Characteristics of effective monitoring systems

There appeared to be rather strong consensus among the 24 respondents about the effects and side effects that might be expected from implementing a system to monitor student learning outcomes in the Flemish education context. Only in a few cases were assumed (side) effects that had been mentioned in the first round contested by other respondents in the second round (e.g., whether such a system would lead to increased segregation between schools). Expected positive effects at the micro level comprised enriched information for students to get to know their competences and make career choices, increased teacher motivation, improved teaching, an improved basis for within-classroom differentiation, and more reliable information for teachers for self-evaluation. Respondents were also convinced that standardized tests contribute to more reliable assessment of students. As one of them said:

> *"So there are good teachers who will feel highly appreciated and will see good scores year after year, and so their motivation will increase. Teachers that are less good, they will see where things are going wrong and they will get better coaching and feel upward pressure. So you'll get an improvement."* (experts, respondent 2, second round interview)

At the meso (school) level, a monitoring system would contribute to the internal quality assurance of schools, but could also be used for school accountability. At the macro level, it is expected that a monitoring system would provide policy makers with a more comprehensive view of the quality of education.

There was also a consensus about several negative side effects, such as teaching to the test:

> *"I'm telling you, we wouldn't have time anymore to work for a week on a project about Bangladesh, or organize a flash mob, or sing a song together. It would mean that our education would become weaker, and that's my major fear."* (schools, respondent 1, second round interview)

In addition, respondents feared a negative emotional impact on students, decreased attention for non-measurable aspects of education, increased competition between schools, schools becoming more selective, decreased school autonomy, and an excessive focus on tests: *"An obsessive culture of figures and data, that doesn't yet exist in Flanders. And well, I wouldn't like it to either."* (experts, respondent 4, second round interview).

Despite the apparent consensus on effects and side effects, respondents strongly disagreed on whether or not a monitoring system should be implemented and how such a system should be conceptualized. This disagreement was based on different values attached to the (side) effects: while some respondents thought that positive effects outweighed negative side effects, other respondents had the opposite opinion.

A total of 21 of 24 respondents thought that the Flemish education system would benefit from some kind of monitoring system. For 20 respondents, the potential contribution to schools' self-evaluation was one of the major reasons, although 18 of these 20 respondents thought that goals at the micro level (individual teachers and students) should also be put forward. A further 8 respondents added purposes at the macro level (for the benefit of policy makers), but no one felt that this should be the primary aim.

In the view of 20 respondents, the implementation of a monitoring system fits into a trust-based perspective, whereas it was only 1 respondent's view that the monitoring system should serve (mainly) the purpose to control quality in schools. A total of 10 respondents added some accountability purposes to the development goals, but the latter took a more central place in their view. Consequently, a monitoring system was seen rather as a mechanism for strengthening trust in schools than as a tool to exercise control. One of the respondents asked:

> *"How do we inspire teachers to improve their own professionalism? In my experience, (…) not by getting tough with them about unachieved results. You need to convey the message one way or another, but it works well primarily by supporting them in their development."* (policy makers, respondent 3, second round interview)

Asked about how to obtain these development-oriented effects, 16 respondents stated that a monitoring system should provide information on students' learning progress throughout their study career. Learning progress is a term used to describe how much a student has learned in a given amount of time. Typically two or more tests are taken from students at different times in order to track their progress. A respondent claimed that, "*There is no benefit for me when my students are compared with the rest of Flanders. (…) It's not interesting for me. I'd rather to know: 'Hey, this student has made great progress.'*" (schools, respondent 6, second round interview). In addition, 17 respondents noted that a monitoring system should deliver added-value information at the school level by comparing the average student results (or progress) within a school with the results of students in schools with a comparable student intake. It was important to these respondents that both student progress and added-value information be considered from a development-oriented perspective, and not as the basis for making (summative) assessments of students or schools. Once again, the trust perspective is preferred over the control perspective.

We need to stress that most of the 21 respondents in favor of implementing a standardized monitoring system gave their personal approval only in the case of several conditions holding true. In addition to the aforementioned possibility to provide a view of student progress, such conditions included broad assessment (a monitoring system not limited to those areas that are easy to test), curriculum-based tests, tests that provide rapid feedback to teachers, results not being made available to the public (as well as other measures that prevent use for accountability purposes), major investments in schools' capacity to interpret and use the results of standardized tests, and a dramatic change in the mindsets of teachers, principals, and policy makers (to output-based thinking).

Finally, 3 out of 24 respondents thought that there is no need to implement such a system. These three respondents (all in the education providers category) thought that the side effects outweighed any potential benefits, as they argued that standardized tests are not in line with the principle of freedom of education and the tradition of school autonomy. These three respondents also feared that—even if it started for development-oriented purposes—any standardized monitoring system would eventually be used to increase authorities' control over schools.

Building scenarios

Several key elements could be deduced from the data collection in the first two research rounds. We discerned two primary key elements: a) the aim of the monitoring system (either trust- or control-based; and oriented at either the meso or micro level), and b) the obligation for schools to use the monitoring system.

Figure 1 presents the eight scenarios discerned based on these primary key elements. Scenario 8 is not included in this figure, as it describes a situation in which no monitoring system would be developed. There is no scenario with a control-oriented monitoring system in combination with optional use, as the nature of control and accountability makes the use of a monitoring system obligatory.

<table>
<tr><th></th><th>Control-based
Obligatory</th><th>Trust-based
Obligatory</th><th>Trust-based
Optional</th></tr>
<tr><td rowspan="2">Micro</td><td>Sc1. Exit exams</td><td rowspan="2">Sc2. Adaptive learning progress monitoring</td><td rowspan="2">Sc3. Test bank for pupils</td></tr>
<tr><td rowspan="2">Sc4. Exit exams with school ranking</td></tr>
<tr><td rowspan="2">Meso</td><td rowspan="2">Sc5. Public evaluation of school quality</td><td rowspan="2">Sc6. Non-public evaluation of school quality</td><td rowspan="2">Sc7. Test bank for schools</td></tr>
</table>

Figure 1. Scenarios 1 to 7

Other key elements included frequency of testing, public availability of results, the possibility to retrieve information about student progress and/or added value for schools, test format (e.g., written, online, adaptive), links with the curriculum, content, and the method of development and correction. The way the primary key elements are conceptualized has an impact on the interpretation of these secondary key elements. We maximized the variation of these elements throughout scenarios 1 to 7. A number of other secondary key elements were identical in all scenarios, as there was already a consensus about them after the first research round, such as the required need for investments into the professional development of schools and teachers. We added an estimate of the financial cost to each scenario. We provide additional explanation regarding each scenario in the next section in discussing the assessment of each scenario by the panel of respondents. A summary of each scenario is added as Appendix 1 to this article.

Evaluating desirability

This final round consisted of a score given by each respondent regarding the desirability of each scenario. We first calculated the mean score per respondent category. Table 2 lists the means of these category means as well as the standard deviations for respondent categories.

Table 2
Scores for scenario desirability (ranging from 1 [minimum] to 5 [maximum])

Scenario		Mean	Sd
1	Exit exams	2.22	0.99
2	Adaptive learning progress monitoring	3.83	0.52
3	Test bank for students	3.56	0.34
4	Exit exams with school rankings	1.53	0.48
5	Public evaluation of school quality	1.71	0.30
6	Non-public evaluation of school quality	3.43	0.62
7	Test bank for schools	3.51	0.61
8	Status quo	2.98	1.18

The four scenarios with a mean score above 3 (neutral) on desirability each describe a monitoring system with a trust-oriented perspective. Generally, the expert panel feared that control-oriented elements would lead to undesirable strategic behavior among teachers, such as teaching to the test. The other primary key elements (micro/meso; obligation) were not deciding factors in judgments regarding desirability.

The two scenarios that included student progress measurements from a development perspective (Scenario 2 and 6) were evaluated as the most desirable, often based mainly on the opportunity to draw conclusions on student progress. As discussed below, in particular the adaptive computer system described in Scenario 2 gave rise to enthusiastic responses.

Scenario 1 included the development of high-stakes exit exams at four key stages during students' school careers: the level of stakes increased at each key stage from 0% to 50% of the decision regarding a student's success. Subsequent scores would also enable measuring students' learning progress. Scenario 4 differs from Scenario 1 as the stakes are high not only at the student level, but also at the school level: in Scenario 4 information about the proportion of students meeting attainment targets is publicly available, paving the way for school rankings. Although respondents were generally in favor of the idea of more objective student assessment, the overall assessment of the desirability of Scenario 1 was negative for 14 out of 24 respondents, due to the risk of teaching to the test and the perceived threat to school

autonomy. Scenario 4 was even less desirable; 2 respondents were in favor of this scenario as it would provide parents with a transparent view of educational quality at different schools, but there were generally strong protests against the idea of public rankings of schools. One of the respondents objected against:

> *"Particularly that public aspect. For me, that's the culmination of an achievement-oriented society and competition and targeting and stress and pressure, and I really don't want that in a primary school. Not for teachers, nor for the children."* (schools, respondent 4, third round interview)

Scenario 2 consisted of an adaptive computer system to monitor student progress in a number of essential skills. At least once per year, every student would be tested on numeracy, literacy, and writing skills. This scenario was assessed by 17 respondents as (very) desirable, and by only 3 respondents as not desirable. Such positive assessments are due to the scenario's focus on progress, which may be beneficial for both individual students and schools. The adaptive nature of this monitoring system was also praised. Such an adaptive system ensures that the test fits neatly with the student's level of competence. Criticism related to the lack of explicit links between the skills assessed and the official curriculum (attainment targets).

In both Scenario 3 and Scenario 7, the authorities developed a test bank containing several reliable tests. In Scenario 3, these tests focused on detailed parts of the curriculum (e.g., fractions) to support teachers' assessment of students. The test bank in Scenario 7 contained tests with a broader scope, so that schools would get a general view of the quality of their education in one of the main areas of education (e.g., mathematics). About half of the respondents judged these scenarios as desirable (13 and 12 respondents for scenarios 3 and 7, respectively, while only 1 and 4 respondents, respectively, assessed these scenarios as not desirable). Such positive assessments are based on the idea that these scenarios provide teachers and schools with reliable information supporting their instruction/internal quality assurance: it is believed that teacher-made assessments often lack reliability or validity. The absence of progress monitoring and teachers' autonomy to use or neglect the tests (and test results) were issues mentioned by opponents of these scenarios. Several respondents also critically reflected on the question of whether or not it is the governments' role to develop a test bank.

The idea behind scenarios 5 and 6 was to support schools' internal quality assurance by providing valid and reliable tests to be taken at regular intervals in order to get a view of students' learning progress throughout their school careers. This idea received broad support from respondents. Scenario 6 was assessed as desirable by 14 respondents, as learning progress has become a central theme in schools' internal quality assurance, and disfavored by only 3 respondents. In contrast, Scenario 5 was unacceptable for 20 respondents,

due to the control-oriented purpose that was added to it: while in Scenario 6 the information on learning output and progress was available only to the school and the students, in Scenario 5 the average student learning progress per school was open to the public.

The average desirability of Scenario 8 (no monitoring system but increased investments into professional development) is harder to interpret, as there was a great deal of variation in respondent opinions. Primarily policy makers and respondents from the schools category thought that maintaining the current system was not sufficient to prepare schools for the future, while education providers were in favor of the status quo. Respondents from the latter category were convinced that negative side effects would outweigh the positive impact.

Conclusion and discussion

The implementation of a system that monitors student learning outcomes would have a major impact on different aspects of any education system. A monitoring system with standardized tests is often considered to be a tool for strengthening authorities' control over schools, but can also be regarded as a trust-based mechanism to support schools and teachers in providing quality education. Whether, and under what conditions, to implement such a monitoring system is a current policy discussion in Flanders. In this study, we have identified a number of key issues that need to be taken into account if Flemish education policy is to consider implementing such a monitoring system. The study has revealed that the focus on strengthening trust should be stronger than the focus on control.

In order for a monitoring system with standardized tests to be effective, a broad base of social support is required among teachers, principals, education providers, and even students (Saunders, 2000); we encountered a strong belief also in Flanders that when a monitoring system is perceived as merely serving policy makers' aims, it is unlikely that its results will be used effectively by teachers.

This study identified several key conditions to allow this basis of support to grow. First, it is important to properly set the goal of the monitoring system. The assessment of policy scenarios shows that a monitoring system should have as its primary aim strengthening trust in schools (even when this would restrict schools' autonomy). Stakeholders think that the education system would not benefit from a high-stakes monitoring system, such as those that are in common use in England and several US states. These high-stakes monitoring systems mainly flourish in contexts with a large degree of competition between schools (Sahlberg, 2011), which is not the case in Flanders.

Earlier studies have shown that teachers value having their assessments confirmed by standardized test results (Beaver & Weinbaum, 2015; Wikeley, 1998). In contrast, the results of this study indicate that this kind of confirmation is not considered to be a sufficient outcome. A monitoring system should lead to greater insight into students' abilities, an expectation that can be addressed by providing teachers with a reliable view of students' learning progress. This entails having several tests taken at different times and the results being linked to one another rather than merely providing one-off measurements. Therefore, the second condition to obtain the required social support is to ensure that the monitoring system enables conclusions to be drawn on the progress that students are making throughout their study career. Although progress monitoring is not commonly used in all monitoring systems, this recommendation has been discussed in earlier debates (Loeb, 2013; Popham, 2005).

Insight into added value for schools was also considered useful by most respondents, but the impact from this factor on the desirability scores of the different scenarios was lower than the impact of monitoring student learning progress. For both learning progress and added value, the results indicate that the ease of drawing practical conclusions was considered to be more important than the technical accuracy of the measurement.

Public availability of standardized test outcomes at the school level has the benefit of being transparent and may increase schools' efforts to provide quality education (Burgess, Propper, & Wilson, 2002; Hoxby, 2003; Jürges & Richter, et al., 2005). However, one potential drawback is that the inevitable school rankings lead to teaching to the test, increased negative stress, and increased segregation between schools (Elstad, 2009; Horn, 2005; Perryman et al., 2011). The high stakes attached to these standardized tests make teachers and principals vulnerable to strategic undesirable behavior and may even induce them to downright cheating during the tests (Amrein-Beardsley et al., 2010). According to Flemish stakeholders, public availability needs to be avoided in order to make any monitoring system generally acceptable. This third condition is required to allay fears about the side effects of publication.

A fourth condition relates to the concern that schools' policy making capacities are currently not yet sufficiently developed to ensure that the implementation of a monitoring system will strengthen the processes of internal quality assurance in schools, which corroborates findings from earlier studies (Onderwijsinspectie, 2013; Vanhoof, Van Petegem, Verhoeven, & Buvens, 2009). There is a need for long-term professional development and coaching of schools with regard to data literacy, which is in line with advice provided by international research on the effectiveness of standardized monitoring systems (Tymms, 1997; Wiliam, 2010).

This Delphi study showed that the issue of monitoring students' learning outcomes and progress is not unambiguous. All too often, the discussion is limited by references to extreme side effects, as observed in some countries with high-stakes testing. From an academic perspective, simple responses to a complex issue should be avoided. The results of this study are an appeal to policy makers and stakeholders for a profound and open discussion on what kind of (trust-based) monitoring system would be most effective for the Flemish education context.

The nature and purpose of monitoring systems varies across educations systems, reflecting values and national traditions with regard to the tension between trust and control (OECD, 2013). Notwithstanding the fact that the results are embedded within the Flemish educational context, the arguments mentioned by stakeholders, as well as the issues that have been identified, may also apply to several other educational contexts. This does not mean that an overall "best practice"' for monitoring systems in order to find the optimal balance between school autonomy and control can be deduced from this study. As we have shown, the value attached to each of the effects and side effects is by nature context-specific and may differ in each education context with its own traditions and aims.

Finally, in our opinion a finding at least as important as the aforementioned conclusions is the experience about the value of the Delphi method for these kinds of policy issues regarding control versus trust in education. This method has proven to be valuable in cases where diverse opinions need to be explored in order to come to a policy decision with the utmost possible consensus.

References

Adler, M., & Ziglio, E. (1996). *Gazing into the oracle: The Delphi method and its application to social policy and public health*. London: Jessica Kingsley Publishers.

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 20–34.

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Educational Policy Analysis Archives, 18*(14), 1–36.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267.

Baird, J., Ahmed, A., Hopfenbeck, T., Brown, C., & Elliott, V. (2013). *Research evidence relating to proposals for reform of the GCSE*. Oxford, UK: Oxford University Centre for Educational Assessment.

Beaver, J. K., & Weinbaum, E. H. (2015). State test data and school improvement efforts. *Educational Policy, 29*(3), 478–503.

Bishop, J. (1998). The effect of curriculum-based external exit exams on student achievement. *Journal of Economic Education, 29*(2), 172–182.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451–469.

Brennan, R., Kim, J., Wenz-Gross, M., & Siperstein, G. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review, 71*(2), 173–217.

Burgess, S., Propper, C., & Wilson, D. (2002). *Does performance monitoring work?* A review of evidence from the UK public sector, excluding health care CMPO Working Paper Series.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–331.

Chong, H., Adnan, H., & Zin, R. M. (2012). A feasible means of methodological advance from Delphi methods: A case study. *International Journal of Academic Research, 4*(2), 247–253.

Cizek, G. J. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. P. Phelps (Ed.), *Defending standardised testing* (pp. 23-54). London: Lawrence Erlbaum Associates Publishers.

Cobb, F., & Russell, N. M. (2014). Meritocracy or complexity: Problematizing racial disparities in mathematics assessment within the context of curricular structures, practices, and discourse. *Journal of Education Policy, 30*(5), 631–649.

Collins, S., Reiss, M., & Stobart, G. (2010). What happens when high-stakes testing stops? Teachers' perceptions of the impact of compulsory national testing in science of 11-year-olds in England and its abolition in Wales. *Assessment in Education: Principles, Policy & Practice, 17*(3), 273–286.

Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science, 9*(3), 458–467.

Day, J., & Bobeva, M. (2005). A generic toolkit for the successful management of Delphi studies. *Electronic Journal of Business Research Methods, 3*(2), 103–116.

De Lange, M., & Dronkers, J. (2007). *Hoe gelijkwaardig blijft het eindexamen tussen scholen? Discrepanties tussen de cijfers voor het schoolonderzoek en het centraal examen in het voortgezet onderwijs tussen 1998 en 2005.* Nijmegen: Netherlands.

Elstad, E. (2009). Schools which are named, shamed and blamed by the media: School accountability in Norway. *Educational Assessment Evaluation and Accountability, 21*(2), 173–189.

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 105–118). London: Springer.

Goodman, D., & Hambleton, R. K. (2005). Some misconceptions about large-scale educational assessments. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 91–110). Mahaj, NJ/ London: Lawrence Erlbaum Associates Publishers.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives, 8*(41), 1–20.

Horn, C. (2005). Standardised assessments and the flow of students into the college admission pool. *Educational Policy, 19*(2), 331–348.

Hoxby, C. M. (2003). School choice and school competition: Evidence from the United States. *Swedish Economic Policy Review, 10*(1), 11–67.

Hsu, C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation, 12*(10), 1–8.

Janssens, F. J. G., Rekers-Mombarg, L., & Lacor, E. (2014). *Leerwinst en toegevoegde waarde in het primair onderwijs*. Den Haag: Ministerie van Onderwijs, Cultuur en Wetenschap; Inspectie van het Onderwijs; Rijksniversiteit Groningen; CED Groep; Universiteit Twente; CITO.

Jürges, H., Büchel, F., & Schneider, K. (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association, 3*(5), 1134–1155.

Jürges, H., Richter, W. F., & Schneider, K. (2005). Teacher quality and incentives: Theoretical and empirical effects of standards on teacher quality. *FinanzArchiv / Public Finance Analysis, 61*(3), 298–326.

Jürges, H., & Schneider, K. (2010). Central exit examinations increase performance... but take the fun out of mathematics. *Journal of Population Economics, 23*(2), 497–517.

Jürges, H., Schneider, K., Senkbeil, M., & Carstensen, C. H. (2009). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review, 31*(1), 56–65.

Karsten, S., Visscher, A., & De Jong, T. (2001). Another side to the coin: The unintended effects of the publication of school performance data in England and France. *Comparative Education, 37*(2), 231–242.

Keeves, J. P., Hungi, N., & Afrassa, T. (2005). Measuring value added effects across schools: Should schools be compared in performance? *Studies in Educational Evaluation, 31*(2-3), 247–266.

Klein, E. D., & Van Ackeren, I. (2012). Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels. *Studies in Educational Evaluation, 37*(4), 180–188.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Educational Policy Analysis Archives, 8*(49), 1–22.

Klenowski, V. (2014). Towards fairer assessment. *Australian Education Research, 41*(4), 445–470.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method. Techniques and applications*. London, Amsterdam, Ontario, Sydney, Tokyo: Addison-Wesley Publishing Company.

Loeb, S. (2013). How can value-added measures be used for teacher improvement? In Carnegie Knowledge Network (Ed.), *What we know series: Value-added methods and applications*. Stanford, CA: Carnegie Knowledge Network.

Marlow, R., Norwich, B., Ukoumunne, O. C., Hansford, L., Sharkey, S., & Ford, T. (2014). A comparison of teacher assessment (APP) with standardised tests in primary literacy and numeracy. *Assessment in Education: Principles, Policy & Practice, 21*(4), 412–426.

Mietzner, D., & Reger, G. (2005). Advantages and disadvantages of scenario approaches for strategic foresight. *International Journal of Technology Intelligence and Planning, 1*(2), 220–239.

Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the university entrance qualification in Germany. *Studies in Educational Evaluation, 37*(4), 206–217.

OECD. (2013). *Synergies for better learning. An international perspective on evaluation and assessment*. Paris: OECD Publishing.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management, 42*(1), 15–29.

Onderwijsinspectie. (2013). *Onderwijsspiegel 2013* [Education mirror 2013]. Brussel: Onderwijsinspectie / Vlaams Ministerie van Onderwijs en Vorming.

Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the pressure cooker – school league tables and English and mathematics teachers' responses to accountability in a results-driven era. *British Journal of Educational Studies, 59*(2), 179–195.

Phelps, R. P. (Ed.) (2005). *Defending standardized testing*. London: Lawrence Erlbaum Associates Publishers.

Popham, W. J. (2005). Can growth ever be beside the point? *Educational Leadership, 63*(3), 83–84.

Ramsteck, C., Muslic, B., Graf, T., Maier, U., & Kuper, H. (2015). Data-based school improvement: The role of principals and school supervisory authorities within the context of low-stakes mandatory proficiency testing in four German states. *International Journal of Educational Management, 29*(6), 766–789.

Ritchie, J., & Spencer, L. (1993). Qualitative data analysis for applied policy research. In A. Bryman & R. Burgess (Eds.), *Analysing qualitative data* (pp. 173–194). London: Routledge.

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York: Teachers College Press.

Saunders, L. (1999). A brief history of educational 'value added': How did we get to where we are? *School Effectiveness and School Improvement, 10*(2), 233–256.

Saunders, L. (2000). Understanding schools' use of 'value added' data: The psychology and sociology of numbers. *Research Papers in Education, 15*(3), 241–258.

Schildkamp, K., Rekers-Mombarg, L., & Harms, T. J. (2012). Student group differences in examination results and utilization for policy and school development. *School Effectiveness and School Improvement, 23*(2), 229–255.

Schwartz, P. (1991). *The art of the long view*. London: Century Business.

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N., & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*(5), 489–499.

Shepard, L. A. (1992). *Will standardised test improve student learning?* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Shewbridge, C., Hulshof, M., Nusche, D., & Stoll, L. (2011). School evaluation in the Flemish community of Belgium. In OECD (Ed.), *OECD reviews of evaluation and assessment in education*. Paris: OECD Publishing.

Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 111–121). Mahaj, NJ/London: Lawrence Erlbaum Associates Publishers.

Smagorinsky, P., Lakly, A., & Johnson, T. S. (2002). Acquiescence, accommodation, and resistance in learning to teach within a prescribed curriculum. *English Education, 34*(3), 187–213.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: SAGE.

Tymms, P. (1997). *Responses of headteachers to value-added and the impact of feedback*. London: School Curriculum and Assessment Authority.

Van Ackeren, I., Block, R., Klein, E. D., & Kühn, S. M. (2012). The impact of statewide exit exams: A descriptive case study of three German states with differing low stakes exam regimes. *Education Policy Analysis Archives, 20*(8), 1–32.

Vanhoof, J., Van Petegem, P., Verhoeven, J., & Buvens, I. (2009). Linking the policymaking capacities of schools and the quality of school self-evaluations: The view of school leaders. *Educational Management Administration & Leadership, 37*(5), 667–686.

Vanhoof, J., & Van Petegem, P. (2007). Matching internal and external evaluation in an era of accountability and school development: lessons from a Flemish perspective. *Studies in Educational Evaluation, 33*(2), 101–119.

Wang, L., Beckett, G. H., & Brown, L. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education, 19*(4), 305–328.

Wikeley, F. (1998). Dissemination of research as a tool for school improvement? *School Leadership & Management, 18*(1), 59–73.

Wiliam, D. (2010). Standardised testing and school accountability. *Educational Psychologist, 45*(2), 107–122.

Wössmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics, 13*(2), 143–169.

Corresponding authors
Maarten Penninckx
University of Antwerp, Edubron Research Unit, Antwerp, Belgium
E-mail: Maarten.penninckx@uantwerp.be

Amy Quintelier
University of Antwerp, Edubron Research Unit, Antwerp, Belgium
E-mail: Amy.quintelier@uantwerp.be

Jan Vanhoof
University of Antwerp, Edubron Research Unit, Antwerp, Belgium
E-mail: Jan.vanhoof@uantwerp.be

Sven De Maeyer
University of Antwerp, Edubron Research Unit, Antwerp, Belgium
E-mail: Sven.demaeyer@uantwerp.be

Peter Van Petegem
University of Antwerp, Edubron Research Unit, Antwerp, Belgium
E-mail: Peter.vanpetegem@uantwerp.be

Appendix 1

Description of the eight scenarios

This overview is only a short summary of the scenarios. The full version of the scenarios evaluated by the respondents was 3–4 pages long.

Scenario 1. Exit exams. Written standardized exams in different courses are taken at the end of primary education and at the end of every second year in secondary education. The tests are taken by all students at the same time. The tests are renewed each year. In primary education, the exam results may affect decisions on students' final grades; in secondary education, the exams results have an increasing weight on students' final grades (15%/30%/50% in the second, fourth, and final grades, respectively; impacts may be positive or negative). Students' learning progress can be calculated based on results from consecutive tests. Information on the results is not made public.

Scenario 2. Adaptive learning progress monitoring. A set of consecutive tests is developed and offered to schools aiming to provide teachers with reliable insight into the development of each student's learning progress with regard to numeracy, literacy, and writing skills at primary schools and numeracy skills, reading and writing skills in Dutch/French/English, and scientific literacy at secondary schools. Tests are adaptive: based on whether their answers are correct or wrong, students get a more difficult or easier set of follow-up items. Tests are taken at least once per year by each student, but schools may decide to give them more often. The results are only used for schools' internal purposes.

Scenario 3. Test bank for students. In this scenario, a test bank is developed. Schools can extract tests (both written and digital) as a support tool to evaluate the extent to which a group of students (or an individual student) has mastered a specific set of attainment targets. Tests are available for a wide array of courses. The tests are therefore very detailed. The test bank also includes tests for socio-emotional aspects as well as diagnostic tests (e.g., for dyslexia). Schools are not obliged to use these tests. The results are only used for schools' internal purposes.

Scenario 4. Exit exams with school rankings. Standardized exams in different courses are taken at the end of primary education and at the end of secondary education. The tests are taken by each student at the same time. The tests are renewed each year. In primary education, exam results determine 50% of students' final grades; in secondary education, the exams result in a binding decision for each student on whether or not a diploma is awarded. Information made public at the school level includes the proportion of students that obtained the attainment targets for every course tested in a school.

Scenario 5. Public evaluation of school quality. Every second year of primary and secondary education, students take written tests in different courses. The tests are obligatory and are taken by all students at the same time. The tests are renewed each year. Individual students' learning progress is calculated based on the results of consecutive tests. The numbers of students with strong/average/poor progress lead to a categorization of schools, which is made public.

Scenario 6. Non-public evaluation of school quality. Digital tests in different courses are obligatory for each student (every second year in secondary education). The results of the standardized tests aim to contribute to schools' internal quality assurance. Different tests are interrelated, enabling mapping of individual students' learning progress as well as the added value for schools based on learning progress. The results are not made public.

Scenario 7. Test bank for schools. In this scenario, a test bank is developed. Schools can extract tests (both written and digital) as a support tool to evaluate the extent to which their education is leading students to effectively meet attainment targets. The test bask includes tests in such areas as social competences and well-being. Schools are not obliged to use these tests. The results are only used for schools' internal purposes.

Scenario 8. No monitoring system. This scenario is limited to increased investment into professional development aimed at strengthening schools to make better use of the data currently available.

Appendix 2

Examples from written questionnaire (round 1) and both interviews (rounds 2 and 3)

Benefits and disadvantages at the micro level

"Students could use these results to obtain a more objective view of their learning results and progress and use this as extra information in addition to other kinds of evaluations. It is independent from the evaluations made by the school. It may help them, for example regarding their decisions regarding higher education tracks. Although it may have benefits, on the other hand there is a risk that this information will be used by the school committee that decides on students' careers (and retention) as students will challenge the committee's decisions with this information in mind. It would also strengthen a kind of 'distrust' in schools' ability to evaluate." (education providers, respondent 4, written questionnaire)

"*I think that there are hardly any advantages from standardized tests for students. (…) Children experience an atrocious amount of stress and they often do much worse on standardized tests than their actual capacity would allow them to in regular circumstances. Questions are formulated in a different way than what students are familiar with and the layout is different (e.g., small boxes for doing calculations).*" (schools, respondent 6, written questionnaire)

"*I like very much that [Scenario 2] fits with the essence of education, the questions that teachers and schools should deal with, namely: how are my students progressing? About the strong students, are we challenging them enough? And about the weaker students, can we help them make more progress?*" (education providers, respondent 4, third round interview)

"*I think it is enormously interesting for a student to know: how am I doing compared to all these other students? That might be motivating as well!*" (education provider, respondent 4, second round interview)

"*The main advantage [of Scenario 1] is that one can evaluate all students equally.*" (policy makers, respondent 1, third round interview)

Benefits and disadvantages at the meso level

"*Teachers and schools receive a more transparent and to-the-point view of their own functioning through the view of learning progress that results from the tests. This leads to teachers taking a more differentiated approach in their education processes. It puts student results into perspective through the benchmark that is provided by the standardized tests. It is a meaningful addition to evaluation data that schools already have available regarding output-oriented functioning.*" (schools, respondent 3, written questionnaire)

"*It may result in strategic behavior by schools, for example by focusing on tests, or by tending towards low scores in prior measurements (for example, for student progress monitoring). It may even further increase competition between schools (which is currently already very high).*" (policy makers, respondent 1, written questionnaire)

"*I see it primarily in light of schools' internal quality assurance and student counseling. Those two aspects. As for the internal quality assurance, I think it is good that a school gets different views of its own quality. Standardized tests may be a tool to provide such a view, on the condition that there are benchmarks, that schools can mirror themselves to certain benchmarks.*" (education provider, respondent 1, second round interview)

"*What I like about [Scenario 6] is that learning progress says something about the school level, and not about the student level.*" (education provider, respondent 3, third round interview)

"*So I asked myself, as a principal would I be inclined in this situation [in Scenario 6] to cheat? To me, teaching to the test is cheating. Justified cheating, as everyone does it. But in this situation, am I inclined to cheat? No!*" (schools, respondent 1, third round interview)

Benefits and disadvantages at the macro level

"*It will lead to comparisons and rankings (good and bad schools, good and bad teachers, students with great or small potential for progress potential, etc.) – and strengthen the 'achievement mentality' in education rather than taking a critical stance towards it.*" (education provider 6, written questionnaire)

"*The most important asset is that policy makers and other stakeholders will get a 'report on the quality of education' on several occasions, and they will be able to see whether learning*

outcomes and learning progress are increasing or decreasing. In other words, it will function as a barometer. It will say something about the quality of Flemish education in general. And it will also enable monitoring of who is best or least served by the education system. Another asset is that many well-intended claims about the quality of education, ones that have sometimes heated the debate in the public arena, will be evaluated effectively, and so the entire debate about this quality will be based on more objective data." (experts, respondent 6, written questionnaire)

"[*Scenario 1] says that the results won't be made public. I would love to believe that, but I've got my doubts about it, about whether schools that receive good scores will keep it to themselves. I think that schools will feel they need to use their scores to show their strengths, which increases the pressure on other schools to make their results public as well.*" (policy makers, respondent 3, third round interview)

"*But regarding this public reporting [in Scenario 4], I think that particularly at the in-between level the factor between publication and the public, namely the journalists and newspapers, will never make a correct interpretation of these data.*" (experts, respondent 1, third round interview)

"*I am charmed by [Scenario 6] as it gets right in between on the one hand being a powerful instrument related to quality assurance and on the other hand leaving schools with sufficient responsibility, sufficient autonomy, and sufficient pedagogical liberty.*" (policy makers, respondent 3, third round interview)