Studies in Educational Evaluation 37 (2011) 180–188

Contents lists available at SciVerse ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels

Esther Dominique Klein *, Isabell van Ackeren

University of Duisburg-Essen, Faculty of Educational Sciences, Institute of Education, Berliner Platz 6-8, 45127 Essen, Germany

ARTICLE INFO

Article history:
Received 1 March 2011
Received in revised form 17 December 2011
Accepted 15 January 2012
Available online 10 February 2012

Keywords:
Statewide exit examinations
Educational governance
Education policy
School quality
School culture

ABSTRACT

Statewide exit examinations play an important role in discussions on school effectiveness. Referring to educational governance concepts, this paper presumes a relation between varying organizational structures of statewide examinations across states, and heterogeneous effects on school actors. It is assumed that their ability to affect work in schools depends on how standardized their procedures are. Therefore, their structural elements must be juxtaposed, and their standardization level must be identified. This paper describes the results of a comparative review of statewide exit examinations using the example of 16 OECD-states alongside the categories Historical Context, Organizational Framework, Marking, and Handling of Results, and discusses the examinations regarding their overall standardization. The study is a basis for further research into how differently structured exit examinations affect school work.

Statewide exit examinations are considered a "powerful instrument for moving schooling in a desired direction" (Noah & Eckstein, 1992, ¶ 2) by holding schools accountable for the achievements of their students (output), and thus enforcing them to align their work with the goals necessitated in the examinations (cf. e.g. Eckstein & Noah, 1993; Keeves, 1994).

In many countries around the world students have to pass exit examinations to graduate from upper secondary education. These examinations assess achievements relative to external standards, and can either be school-based or statewide. The former are generated in the individual school, mostly by the class teacher, and then usually authorized by the respective authority. Oppositely, the latter are organized by a central entity. This implies that all students within a certain administrative area[1] have to take the same tests, and usually do so at the same time. Teachers, who have a lot of influence on school-based, but only little, if any, influence on statewide examinations, are therefore prompted to adjust their work to the standards implied in them (cf. Bishop, 1998; Wößmann, 2003). Thus, statewide examinations are assumed to affect and regulate work at both the school- and classroom level by setting minimum standards (*qualification*), promoting quick implementation of new syllabi (*innovation*), helping to improve instruction (*professionalization*), and increasing commitment (*motivation*). This is supposed to happen in two ways. On one hand, the examinations provide explicit and transparent information on what competencies students must have acquired by the end of upper secondary education, and on where schools have to improve (*support*). On the other hand, they hold school members accountable for their work, and link the graduation of the students to the examinations (*pressure*). Thus, the examinations shall provide incentives for both educators to elaborate strategies to improve their school and instruction, and students to raise efforts to meet the standards (on accountability tests in general see e.g. Amrein & Berliner, 2002; Vogler, 2006, p. 2f). As a consequence, statewide exit examinations – as a seemingly more standardized instrument than school-based exit examinations – have gained more and more popularity during the past decades: Especially within the context of changed governance structures, increased school autonomy, and decentralized decision powers, statewide exit examinations are seen as an instrument with which the school system can be "navigated". As an example, a number of the German federal states have switched from school-based to statewide *Abitur* examinations after 2005. These changes happened mainly in reaction to the poor PISA results and were explicitly motivated with the expectation that the "new" way of assessing students would lead to the positive effects mentioned before. Comparable tendencies can be observed in the USA (cf. Center on Education Policy (CEP), 2008).

Within political debates, statewide exit examinations are usually treated as unambiguous, and variances are predominantly discussed with regard to the achievement standards that seem to

* Corresponding author. Tel.: +49 201 183 2961; fax: +49 201 183 3093.
E-mail addresses: dominique.klein@uni-due.de (E.D. Klein), Isabell.van-ackeren@uni-due.de (I. van Ackeren).

[1] Despite the inaccuracy of "statewide" – as the administrative area might also be regional – the term will be used owing to a lack of a more accurate alternative.

0191-491X/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.stueduc.2012.01.002

differ between examinations in different states. Against this background, the German Institute for Educational Quality Improvement (IQB)[2] – to give a national example – is commissioned to specify and empirically evaluate national educational standards for statewide examinations to ensure comparability across the German *Länder*. On the first look, the differences in content seem to be the main aspect that leads to variation between the examinations. Upon closer inspection, however, the examinations are noticeably dissimilar with regard to their organizational structures and the standardization of procedures – even between the German *Länder* (cf. Kühn, van Ackeren, Block, & Klein, 2009) which share a widely common political and historical background. The seemingly unambiguous term "statewide exit examination" (or *Zentralabitur* in the German context) describes diverging procedures, and the differences in the design and standardization of statewide examinations are even more distinct when examinations are contrasted in an international comparison.

This leads us to assume that accordingly, the effects of these examinations on personnel in school are divergent as well due to the different structures, and that not all of these structures have the potential to affect schooling in the same way. This assumption is supported by the current research on this topic (see next paragraph). Differing restraints and bonds linked to the varying procedures and standardization levels might affect the actions of local school actors and thus cause heterogeneous effects. Therefore, to thoroughly analyze how statewide examinations actually do affect schooling presupposes knowledge of heterogeneous structures and of indicators of the standardization level (e.g. uniform and explicit criteria, methods, processes, and practices for the regulation of statewide exit examinations). At present, however, a current systematic comparison of examination structures and indicators of standardization to back up research does not exist. To handle this desideratum, the second part of this paper offers a study carried out by the authors, which compiles the current structural varieties of statewide exit examinations at the end of general upper secondary education within 16 OECD-states, with special attention given to their standardization level. The study provides a basis for further research into how differently structured exit examinations affect the work of local actors.

Status of research

Apparently, the easiest way to estimate how examinations affect school quality would be to compare achievements of students in states with or without statewide exit examinations in studies such as PISA and TIMSS. A number of calculations of this kind exist (cf. e.g. Bishop, 1998; Fuchs & Wößmann, 2007; Hanushek & Wößmann, 2007; Jürges, Schneider, & Büchel, 2003; Wößmann, 2003), but their results are inconsistent and vary among different domains and age groups. Apart from that, they also neglect that differently structured examinations on one hand and multiple personal and circumstantial factors on the other hand influence the achievement of the students and how people in schools handle examination requirements. A possible way to take the latter factors into account is to inquire about how the exams are received and reflected on by actors in schools. Research of that kind, especially concerning high stakes tests, exists in great quantities and with inconsistent results (for an overview cf. e.g. Au, 2007; Nichols & Berliner, 2007; Volante, 2007): The findings do not only display the hoped-for positive effects, but also severe side-effects. In several cases the tests promoted the reallocation of teacher aid, teaching to the test, forbidden assistance during examinations, fixation on short-term improvement, and omission of innovative teaching methods. Additionally, teachers reported deprofessionalization, threats to their professional integrity, a decline of trust, and decreasing motivation (cf. e.g. Bellmann & Weiß, 2009, p. 292ff; Hamilton, Stecher, & Klein, 2002, p. 85ff). The tests in question usually had high stakes not only for the students taking the examinations, but also (and very often exclusively) for their teachers and schools. In many cases, however, statewide exit examinations are not coupled with sanctions for schools and teachers directly, so that the findings described above only limitedly apply to statewide examinations.

The scant existing research on statewide exit examinations in particular neither confirms nor confutes the results on high stakes tests. On one hand, positive or neutral effects, with regard to cooperative and supportive structures (e.g. Maag Merki, 2008), for instance, or the use of student-centered instructional methods (e.g. Vogler, 2006; Vogler & Carnes, 2009), have been reported. On the other hand, some authors confirm that teachers reallocate resources concerning contents and aid (e.g. Gillborn & Youdell, 2000). Effects on instruction seem to depend to a large degree on subjects (cf. Maag Merki, Holmeier, Jäger, & Oerke, 2010). Moreover, it becomes apparent that how school actors perceive and act upon the examinations is probably affected or even provoked by collective attitudes (cf. Maag Merki, Klieme, & Holmeier, 2008) and capacities at school level (cf. DeBray, Parson, & Woodworth, 2001).

To sum up, research confirms that statewide tests, and probably statewide examinations, do affect prior instruction and working structures. However, they also show that the very heterogeneous effects are influenced by more than just whether the examinations are statewide or school-based. The examinations represent the attempt of the administration to regulate the work at school and classroom level. Therefore it is necessary to investigate how differently structured and standardized examinations influence learning at school and classroom level and to analyze why people in schools react differently to the examinations. A prerequisite for future research therefore is that examination procedures are investigated in more detail before their effects can be illuminated.

International comparison of statewide exit examinations

This study assumes that if the examination is presumed to regulate work in schools in terms of quality improvement, the standardization level of its procedures, within specific governance structures, is decisive. In this respect, standardization addresses the degree to which the different steps of the examination are taken out of the control of an individual school, which diminishes the influence of individual and organizational factors. For instance, if papers are marked by the individual class teacher, even with regulating marking schemes, the teacher may adapt his or her marking to the way class content has been taught. If, on the other hand, the marking is done by an external marker, and thus has a higher standardization level, the teacher will have to adjust instruction more firmly to the examinations, because he or she has no influence on the marking.

Existing studies investigating how the examinations affect school work, instruction, and learner performance usually do not account for the standardization level of their procedures, their organizational frame, and the interplay of governmental attempts to regulate schools on one hand and the self-regulation of the individual school on the other hand. Nevertheless, for a reliable analysis, the effects of statewide examinations must be examined and discussed against their structure and kind of standardization, and how this affects school actors.

[2] cf. http://www.iqb.hu-berlin.de/bista/abi, 26 February 2011.

Table 1
Indicators and range of standardization levels.

		Lower standardization	Medium standardization	Higher standardization
Organizational Framework	Administrative area	Regional		Statewide
	Type of examination	Written statewide and school-based examinations	Written statewide, oral school-based examinations	Written and oral (practical) statewide examinations
	Number of examinations	1	4–6	All subjects taken
	Ratio statewide to school-based written examinations	School-based outnumber statewide examinations	School-based equal statewide examinations	Statewide examinations only
	Types of subjects	Core subjects only	Core subjects and subjects of specialization	All subjects taken
	Task development	Individual teachers	Central commission without training	Central commission with training
	Candidates' choice: subjects examined	Free choice	Compulsory and optional subjects	All subjects taken
	Candidates' choice: difficulty level	Uncoupled from course level	Dependent on course level	Same level for all students
Marking	Marking stages	1	2	3
	Markers	Internal marker included	External markers from another school	Central commission
	Anonymity of candidates	Not given	Partially given	Completely given
	Marking standards	Defined ex post		Defined ex ante
Handling of Results	Share of statewide examination results in final grade	Less than 30%	50% or more	100%
	Publication of individual school results	Not possible	Possible, but not executed	Possible and executed

If not stated otherwise, the table refers to statewide written exit examinations only.

In this study, the structures of the examinations are reviewed in an exploratory international comparative survey.[3] The survey covers examinations at the end of general upper secondary education (ISCED 3A). The states are chosen upon the following criteria:

(1) they are of a comparable socioeconomic status (member countries of the OECD), so that effects are not distorted by higher poverty, illiteracy, etc.;
(2) they are not organized federally, as the usually diverse graduation procedures of federally organized systems can at this point not be disclosed differentially;
(3) at the end of upper secondary schooling, students take an examination that is completely or partly *statewide* as defined in the beginning.

Accordingly, the following 16 OECD-states have been chosen: Denmark, Finland, France, Greece, Hungary, Ireland, Italy, Luxembourg, Netherlands, New Zealand, Norway, Poland, Portugal, Scotland, Slovakia, and United Kingdom (England, Wales and Northern Ireland).

Methodology

The varieties of examination procedures are outlined in a systematic comparative summary. Data has been collected through content analysis of official and legal documents provided by the countries. The analysis starts with the portrayal of the education systems provided by the information network on education in Europe, Eurydice, and, when attainable, the websites of the respective ministries and examination commissions, and the relevant examination regulations have been taken into account as well. Additionally, in some states experts from the education authorities have been consulted in unstructured interviews. In the analysis, which can only be displayed in extracts here, the state-specific structures and procedures are described, juxtaposed, compared and discussed with regard to their possible relevance for impact analyses.

The analysis is divided into four categories. It first highlights the *Historical and Cultural Context* of the examinations with regard to examination traditions, motives for implementing statewide examinations, and state-specific governance "philosophies". The possible shapes of elements indicating the standardization level are then described chronologically alongside the categories *Organizational Framework, Marking,* and *Handling of Results*. *Organizational Framework* includes information such as the number and share of statewide examinations, how and by whom examination tasks are developed, and whether candidates have a choice in proficiency levels and contents. *Marking* covers the stages of marking, choice of markers, marking standards, and students' anonymity. Finally, *Handling of Results* includes the meaning of the examination results for final grades, how overall results are reported to schools, and whether they are coupled with high stakes for teachers and schools.

Findings: the standardization level as indicator of the examinations' capacity to regulate schooling

Throughout all categories analyzed, the comparison reveals wide variation in the conduct of the examinations. Therefore, common procedures for each element are illustrated and supplemented with further rather highly or lowly standardized varieties, pinpointing the scope of possible standardization levels within each element (see also Table 1). Unless otherwise noted, the information only refers to *statewide* examinations.

Historical and cultural context

Education systems have established statewide exit examinations for differing reasons and within distinct time-frames. In Ireland, the predecessor to today's *Leaving Certificate Examination* was introduced in 1879 to set a general standard of secondary education, and today still mainly serves the purpose of certifying

[3] As of 2006/07; the review served as a preliminary study for an ongoing survey on the differing purposes and effects of state-wide exit examinations in selected countries. The results of the main study cannot be reported yet.

the completion of upper secondary education and allocating graduates (cf. State Examinations Commission (SEC), n.d., p. 34f). The Finnish Ylioppilastutkinto, too, was introduced in the 19th century mainly to regulate the influx into higher education,[4] but nowadays also serves as an instrument that provides data feedback for school improvement and helps monitoring the overall quality of the education system. In France, the *Baccalauréat,* established in 1808, is embedded in a school system that is committed to assuring equal chances, whereas in England, the *A-Levels* are integrated into market-based school improvement (cf. van Ackeren, 2003; Cole & John, 2001) and, in addition to certifying graduation, provide information on the "value" of schools for their "customers". In the Netherlands, the mix of the long standing *centraal examen* and a school-based examination introduced in 1968 is supposed to balance autonomy of schools and external control (cf. Louis & Versloot, 1996). Here, too, the statewide examinations are used to hold schools accountable and inform customers' choices, albeit the consequences are much lower. In some other states, however, statewide exit examinations have only recently been introduced or revised (e.g. Hungary, Slovakia, some German *Länder*) with the explicit goal of improving effectiveness through the control of the outcome (cf. above, and e.g. National Institute for Public Education, 2003).

The different contexts within which the examinations were established as well as the institutional traditions and cultural beliefs about the functioning of schools within different education systems have produced examinations with rather specific functions in each country. Apart from the fact that these functions have in turn brought forth very distinct procedures in the examinations, it is also very likely that at the school level, the perception of and action upon the examinations will be influenced by their particular functions: School actors who believe the examinations are threatening the existence of their school or devaluating their professional expertise will probably react differently to them than school actors who see the examinations as an opportunity to learn and improve, and feel supported by the state through the examinations.

Organizational Framework

Administrative area

When considering who is responsible for the overall organization of the examinations, the majority of states have one central authority, so that students throughout the state take the same examinations within one subject. In the UK, however, although the fundamental structure of the *GCE A-levels* is basically the same across the state, each country has its own qualification authority. The tasks are developed by private, non-profit examination boards (*Awarding Bodies*) between which schools may choose. This regional differentiation, however, must be distinguished from states with regionally independent graduation procedures, as is the case for instance in Canada, Germany, and the USA. A procedure that is mainly in the hand of state departments might possibly be seen as an attempt of the state to maintain control over and standardize the examinations, and through this pre-structure at least the direction of school development processes. The policy of the UK, on the other hand, expresses the principle of improving schools with competitive elements. Schools have to scrutinize the offers of Awarding Bodies, and choose the one that best suits their circumstances. This, however, also caters for increased local and regional differences in the contents of examinations and the information that is being fed back to school, especially since at present, there is no national curriculum for upper secondary education.

[4] http://www.ylioppilastutkinto.fi/en/index.html, 10 August 2011.

Type and number of examinations

Two thirds of the states have statewide examinations only, whereas the others have additional school-based examinations. In these cases, usually the number of school-based examinations equals or outnumbers that of statewide ones. In the Netherlands, for instance, the statewide examinations are accompanied by a coequal *schoolexamen* for each subject. In some subjects, students sit a school exam only.

In most states, only written examinations are carried out on a statewide level (except e.g. Ireland). However, the number of written examinations may vary, depending for instance on subject groups and difficulty levels (see below). In some cases, students have to sit statewide written examinations in most or all of their subjects, although in the UK, as an example, the number of subjects they enroll in – six – is comparatively limited to begin with. In some other states, where written examinations are generally statewide, there is only a limited number of subjects tested in a written examination. In Slovakia, candidates sit only one statewide written examination, whereas all other examinations are school-based oral examinations. Alternatively, Italian students only take two statewide examinations, which are supplemented with two school-based examinations – one written and one oral.

In this respect, how pressurized students feel by the examinations is different, and, mediated through this, the examinations might have a lesser, or at least different, impact on school processes and instruction in countries where students can balance their examination results with internal assessment results. However, this assumable effect will probably be cancelled out if there are additional stakes for schools.

Types and choice of subjects

The selection of examination subjects is, as a matter of fact, associated with the organization of upper secondary schooling. In all states, a statewide examination can be taken in the national language. Additionally, most states examine a certain number of core subjects (if there are any) – mostly foreign languages, and, less frequently, mathematics, humanities and science. Apart from that, the selection of subjects usually depends either on subject groups, specialization, or difficulty levels. If during upper secondary schooling, students select a specific field of specialization, or choose one or more subjects at higher level, they most commonly sit statewide examinations in these subjects and one or more subjects of the common core. In France, for instance, students of the *baccalauréat scientifique* have statewide written examinations in a number of their science subjects – here, they have a partial choice as to which – and in several core subjects like French or mathematics. In Portugal, the common core is more reduced, as students have statewide examinations in Portuguese and three subjects of specialization.

Usually, students may at least partially choose in which subjects they want to sit the examinations. This choice is very open in the UK, as subjects can be studied (and thus assessed) in any combination. In contrast to that, Finnish students have rather low freedom of choice. They take four examinations, with Finnish or Swedish as mother tongue being compulsory for everyone, and three other examinations from a pool of four (the other national language, a foreign language, mathematics, and general studies – here the choice is broader), but may also take further optional examinations.

The conditions of subject choice might possibly lead to very different ways of processing what is required for both students, who can or cannot avoid certain subjects, and schools and teachers. In particular, a very close focus on a handful of core subjects might for instance evoke that resources and attentions are reallocated to these subjects, which would increase the chance for improvement here, but also decrease the chance of other subjects. A very open

subject choice, on the other hand, might induce avoidance, in that schools encourage students to primarily choose subjects in which the results are likely to be good. This might likewise lead to a situation where schools try to maintain their status quo instead of focusing on improvement.

Task development

In the Netherlands and the UK the tasks are developed by independent non-profit organizations. In the UK, the examination tasks (and the syllabi they are based on) are created by *Awarding Bodies*. In the Netherlands, the tasks are designed by the testing institute Cito (*Centraal Instituut voor Toetsontwikkeling*). Cito, in contrast to the task development in the UK, has to stick closely to the syllabi and guidelines provided by the state-run *College voor Examens*.

Apart from these exemptions, developing examination tasks is commonly the responsibility of state-run institutions subordinate to the education ministry. The tasks are usually designed by special commissions of diverse compositions and size. They are predominantly staffed with teachers in France and consist of subject teachers and university experts in Finland. The subject groups of the Dutch Cito are staffed with Cito employees who are psychometric experts, and teachers who work for Cito part-time.

The various group compositions conceivably embody different expectations towards the quality of content and measurement in examination tasks. Thus, for instance, most countries involve teachers in task development to assure that the tasks reflect what happens in the classroom. In addition, sometimes test and statistics experts are included. This might signify that value is set on accurate measurement, which probably is especially important when the examinations have high stakes for schools (e.g. Netherlands, UK). Including experts of the respective subject didactics in turn may indicate that the examinations are supposed to be innovative, and in this carry novelties into classrooms. From a school's perspective, tasks that are secured with expert knowledge probably also limit the opportunity to blame shortcomings in the tasks when results are below expectations.

In almost all states, the tasks are checked for content and form throughout several steps, and in some states, are tested under specific conditions (e.g. France).

Candidates' choices in difficulty level, type and task

In most states, the difficulty level of the examination parallels that of the course the student enrolled in during upper secondary education. If all students take the courses at the same level, they will usually take the examinations at the same level, too (e.g. Netherlands). Sometimes, however, the choice of difficulty level in the examination is unrelated to the level of the course. In Finland, the examinations in mathematics and the language subjects are arranged in two different levels from which candidates may choose despite their course levels, as long as one exam is taken at advanced level. The procedure in the Netherlands reflects the function of the examinations, which is mainly to provide a standardized and comparable measure that schools can use to assess their own situation. At the same time, the *centraal examen* is accompanied by non-standardized school examinations that leave room for more individualized assessment. In contrast, other countries like Finland have systems in which students are supposed to show a certain level of proficiency in some subjects (which in the Netherlands is part of the school examination). The fact that this proficiency must be proven in a statewide exit examination might lead schools to increasingly challenge or support their students.

With regard to type and content, the freedom of the candidates is generally more restricted in most states. Whether they are allowed to choose from one out of two or more different examinations is connected with the subject field in the majority of states. In the humanities, students are more likely to have a choice between different examinations than in more technical subjects like science, math, etc. This may cause diverging responses in different subjects, as teachers might feel less restricted in an examination that leaves room for choice. Albeit, it might also be that teachers in science subjects, as opposed to, for instance, literature, do not feel quite as restricted by decreased choices due to the nature of their subject: Science subjects are generally conceived as being well-defined, with "hard facts" and clearly sequenced contents. An alternative to whole examination papers being exchangeable is that students only have a choice in parts of the examination, which then consists of compulsory and optional components (e.g. in Scotland). In some states, students are not offered a choice in any subject at all (e.g. Netherlands).

Marking

Marking stages

The number of marking stages varies from only one to three obligatory markings. In most Anglophone states, papers are marked in a one-stage process, whereas others generally rely on two independent markings. The latter approach is in some states linked with a possible third marking if the results of the first and second stage are too inconsistent (e.g. Denmark). A regular third marking is undertaken in Luxembourg. The number of stages is usually linked to the characteristics of the markers, and does not necessarily reflect a more or less restricted procedure.

Internal and external markers

Not all of the marking stages are carried out by external markers in all states; instead, in – only a few – states, internal markers, such as the class or course teacher, are involved in the procedure as well, although as a rule, the marking is not carried out by internal markers only. In Finland, the preliminary marking is conducted by the course teacher, the second – and decisive – marking by markers of the state examination board. In the Netherlands, papers are marked by the course teacher as well as another teacher from a different school. Most other states, however, keep the marking external.

External marking, too, can have different shapes. It may be assigned to teachers of other schools, as in France or Luxembourg, or conducted by a higher-level entity completely. In most English-speaking countries, the examinations are marked by persons recruited by the authority, who go through a marker training. In some states, oral and practical examinations are also graded by external markers (e.g. Ireland).

Anonymity of the candidates

The ratio of states in which the markers actually know the identity of the candidates (e.g. Netherlands, UK), and those in which they are anonymous (e.g. France, Luxembourg), is rather balanced. However, the anonymity and the marking procedure seem to be related in that the candidates are more likely to be known if the marking is done by an internal marker. Especially with regard to the objectivity and integrity of the examination results, this aspect seems to be essential. However, results that have been determined without the school having a hand in this may have a double-edged influence on school processes. On one hand, schools, and especially teachers, might be more inclined to believe that an assessment of their students reflects their true abilities, and thus be less prone to shift the blame for bad results to an unfair marking, when they are involved in the marking process. On the other hand, knowing that the examinations will be assessed by an outsider and thus cannot be "straightened out" by the teacher might urge schools to improve much more. However, it might also urge schools to increasingly play the system.

Marking standards

To guarantee that papers are marked in a reliable manner, marking standards are normally defined and marking schemes developed ex ante. Only little information can be gathered with regards to the content of these standards and schemes. It still becomes apparent that the shape of marking schemes responds to the type of tasks, that is, in examinations with closed question formats (multiple choice or short answers), the instructions tend to be more restrictive than in essay type questions. In open questions, usually only potential "expectable" answers and acceptable alternatives as well as marking recommendations are issued. Apart from that, marking schemes usually lay down directions about the awarding of points.

Nevertheless, the marking standards are not inalterable. In fact, in most states, the standards are revised ex post. In England, the marking is followed by a *Standardization Meeting*, in which the thresholds for grading are determined, according to the average achievement in this examination, and with regard to the achievements of the preceding years. This way, the normally criterion-referenced grading of statewide examinations is supplemented with social benchmarks, and the countries maintain the possibility to adjust standards to the current cohort, which is especially important in view of graduation rates.

Handling of Results

Weighting of the final grade

Two states where the final result a student will receive when he or she graduates consists of the statewide examinations' results only are France and Ireland. In most states, however, internal assessment is at least partially considered in the overall qualifications, which leaves room for the schools and teachers to focus individual aspects and respond to their students' interests. The weighting of the different components can differ considerably from one state to another, and there does not seem to be consensus across states. For instance, in Portugal, internal grading counts seventy percent, while examination results count only thirty percent. In the Netherlands, the results of the *centrale examens* and the *schoolexamens* are equal. In contrast to that, in Luxembourg, the statewide examinations count twice as much as internal grading. In some states, students get two different certificates, one containing the results of the examinations, and one stating the results of continuous assessment (e.g. Finland). The value of the latter for the application in further education institutions, however, differs.

Feedback and use of results

In addition to feedback to individual schools, in a number of states a report on average achievements in the examinations is usually published. In Ireland, the *Chief Examiners Report*, published for a number of subjects, provides information for instance on the development of results within the past years and on frequently occurring problems, and gives general recommendations for how students and teachers should handle the examinations, or what aspects they should focus during instruction. The Scottish *Principal Assessor Report* is a comparable document. The feedback on results, together with recommendations for improvement, even though not necessarily tailored to the situation of each school, might possibly be a stronger lever for school development processes than data feedback on its own.

Furthermore, in England, France, and the Netherlands – to give just a few examples – the state publishes the results of individual schools. In the English *League Tables*, the results of schools are compared on a regional and national level as part of a school improvement that relies on quasi-market elements and competition among schools ("naming and shaming"), and combines the performance with stakes for the schools. This system, however, does not only lead to the wished-for improvement efforts, but may also provoke "educational triage" (Gillborn & Youdell, 2000) and attempts to play the system and distort results.

In the Netherlands, the publication of school results was not initially driven by the state, but rather was provoked by the press in the 1990s. As a result, today the inspectorate publishes the *Opbrengstenkaart* ("output card") with information about each school, which includes the results of the *centraal examen*, among other data, and also relates the exam results to the condition of the school. Nonetheless, in some states, the right to publish the results of individual schools is exercised by the media, though not by the government (e.g. Finland). In other states, although often demanded by parents and media, the administration prohibits the publication of individual examination results (e.g. Ireland).

How standardized are statewide exit examinations? Summary and discussion

The preceding chapter portrays how heterogeneously examinations that all qualify as statewide exit examinations can be organized, and thus supports the plausible assumption that the observed heterogeneous effects described in Chapter 2 can – at least in parts – be traced back to heterogeneous examination structures.[5] There are only few elements on which the majority (though not all) of the states share a common design. These similarities are that the examinations usually take place at the same time and date across the state, the national language subject is tested in a statewide examination, and the examinations are marked by external markers.

For most elements, however, differences prevail. These differences can likewise be seen as benchmarks for rating the standardization. The authors refrained from developing an instrument with which a "standardization degree" of the examination procedures could be rated quantitatively and which would "reduce" the various dimensions of the analysis to one "standardization scale". Instead, the elements of difference are used to qualitatively compare the standardization of the examinations. The most pivotal elements where differences become obvious are:

(1) the extent of the statewide examinations, in particular
   - whether all or just selected subjects are assessed, and what the selection of subjects is based on,
   - whether the examinations are only statewide when they are written, or whether this also applies to oral and practical examinations, and
   - whether there are additional school-based (written) examinations,
(2) candidates' choices in subjects, topics and tasks, and difficulty levels;
(3) the marking process, in particular
   - in how many stages and by whom the marking is done,
   - how marking standards are assured, and
   - whether candidates remain anonymous,
(4) the weight and utilization of results, in particular
   - whether internal assessment grades are included in the overall qualification, and what is their share in the final results, and
   - whether the results of individual schools are made public.

[5] It should be emphasized, however, that our analysis is focused on rather formal dimensions of statewide examinations. Furthermore, the ways actors react upon these within different cultural contexts might also be relevant for explaining the differences across various education systems.

Table 2
Examination procedures in selected countries.

	Finland	France	Ireland	Netherlands	UK
Name of examination	Ylioppilastutkinto	Baccalauréat Général	Leaving Certificate Examinations	Eindexamen vwo	GCE A-Levels
Established	1852[a]	1808	1879/1925[b]	1968[c]	1951[d]
Organizational Framework					
Administration	Statewide, exam commission	Statewide, Ministry of Education	Statewide, exam commission	Statewide, exam commission and testing institute	Regional (school choice) *awarding bodies*
No of written exams	Min 4	Min 9	Min 5	≈8	Min 6
Scope	SW: written SB: none	SW: written and oral	SW: written and oral SB: none[e]	SW: written SB: written and oral	SW: written SB: none
Ratio SW to SB	SW only	SW > SB	SW only	SW < SB	SW only
Choice of subjects	Core + choice	Core + specialization	Irish + choice	Core + specialization	English + choice
Task development	Exam commission: teachers and/or university experts	Exam commission: mainly teachers, 1 inspector, 1 university expert	Exam commission: mainly (former) teachers	Exam commission: teachers and psycho-metricians	Exam commission (not specified)
Training	No	(not spec.)	Yes	Yes	(not spec.)
Choice in difficulty levels	Subject-specific, *not* linked to course level	Yes, linked to course level	Yes, *not* linked to course level	No	No
Marking					
Stages	2	2	1	2	1
Marker(s)	Course teacher (preliminary) + marker of exam commission (decisive)	Teacher of another school	Marker of exam commission	Course teacher + teacher of another school	Marker of exam commission
Standards	No[f]	Yes	Yes	Yes	Yes
Anonymity	No	Yes	Yes	No	No
Handling of Results					
Final grade	SW results only, separate certificate for internal assessment	SW results only	SW results only	SW and SB results make 50% of the grade	SW and SB results
School results published?	No	Yes	No	Yes	Yes

Abbreviations: SW, statewide exit examinations; SB, school-based exit examinations; not spec., not specified or no information available.
[a] Introduced as entrance examination to the University of Helsinki; later changed into state examination (no further specifications available).
[b] Examinations were introduced in 1879; after the foundation of the Republic of Ireland, new programs were introduced for the Leaving Certificate Examinations, which were first held in 1925.
[c] Refers to the system of statewide and school-based exams that is used today.
[d] The predecessor of the A-levels, the Higher School Certificate, had been introduced in 1918.
[e] Practical coursework in some subjects: supervised by teachers and marked by marker of exam commission.
[f] Standards for the preliminary marking are developed by subject associations, but are not binding.

Within this scope, some states conduct highly standardized examinations, while others have a lower overall standardization degree. To illustrate how different the examination structures within different contexts are, Table 2 depicts the features of examinations using the examples of Finland, France, Ireland, the Netherlands, and the UK.

The Irish *Leaving Certificate Examination*, as well as the *A-levels* in the UK, show a high standardization level across most elements, as there are no school-based examinations at all, and individual schools have only little influence as to the conduct of the examination: even the marking is done completely by markers in the respective examination commissions. Nevertheless, the UK examinations are slightly less standardized, as they are not prepared by a central entity. What is more, both countries' examinations are less standardized in certain areas, such as what subjects are prescribed for all students. In this respect, the examinations from Finland, France and the Netherlands are more standardized, as they prescribe core subjects, and, in France and the Netherlands, also regulate which subjects of the specialization will be tested. On the other hand, when we look at the marking processes, these three countries are less standardized than Ireland and the UK. In Finland, the decisive marking done by the examination board is not based on a standardized marking scheme. The Dutch examination is rather standardized with regard to the development of papers and marking schemes of the statewide part, which is depicted in Table 2; the marking of the papers, however, is in the hands of the schools, and how schools design their *schoolexamen* is largely open. In addition, since the results of the school examination count for half of the final grade of students, the work of schools might be less influenced by the impact of a statewide exit examination.

Some other states which are not illustrated in Table 2 have a lesser standardized examination procedure altogether. In Slovakia, for instance, only one test is conducted in a statewide manner. Even if the standards in the categories described in Table 2 are very rigorous in this one subject, the overall standardization of the examination is rather low, Additionally, the emphasis of one subject might cause local actors to especially focus on this subject (cf. Maag Merki & Holmeier, 2008). As mentioned, the standardization of current procedures in the German *Länder* differs, but they nonetheless share a medium to low standardization (for a comprehensive overview of the procedures in Germany, see Kühn

et al., 2009). This commonality of all the *Länder* signifies that common political and cultural attitudes towards examinations and school accountability influence how the examinations are organized. The specific intentions and expectations of the administration regarding the regulation of schools must therefore be considered when differently structured examinations and their effects are analyzed.

The data shows that even countries with a comparably high standardization, in our examples in Table 2, this applies to Ireland and the UK, still are considerably different. Ireland and the UK differ, for instance, in the administration of the examinations, the anonymity of candidates, and the handling of results. Accordingly, general standardization patterns can be detected for neither the five countries in Table 2, nor the other countries in the sample, even when they share common governance structures. In states where the examinations are part of a competitive market, for instance, the overall structures of the examinations only have little in common: In England and New Zealand,[6] for example, the examinations are in some parts comparable (e.g. the recruitment of markers). Nonetheless, the structures are dissimilar in a lot of elements, and even with regard to the standardization level of these differently designed elements. This might be because the examinations have been introduced in distinct time frames and, although having the overall goal of improving school effectiveness, include distinct additional regulation intentions of the respective countries.

Comparably, regulation intentions are also reflected in the Dutch examination, which is split into the *centraal examen* and the coequal *schoolexamen* and thus serves as a "compromise" in a school system with highly autonomous schools (cf. Louis & Versloot, 1996, p. 255). The *centraal examen* shows a high standardization degree in most elements of task development (see Table 2), whereas form and contents of the *schoolexamen* are the responsibility of the individual schools. The statewide examination provides a benchmark upon which the performance of students – and schools – is made comparable, whereas the school-based examinations embody the concern for a more reliable overview of students' achievements assessed in a way that is close to how subject matter was taught. Similar tendencies can be detected in other countries not depicted in Table 2. In Portugal, for instance, the results of internal assessments weigh more than twice as much as those of the statewide examination.

The Finnish examination, as a different example, is embedded in a governance structure which is broadly decentralized and has no central school inspection (cf. Laukkanen, 2008): The *Ylioppilastutkinto* is the only nation-wide external test, and it is mainly meant to provide information for schools.

The comparison shows that there is no one-size-fits-all approach to the conduct of statewide exit examinations. Rather than that, the different practices display widely diverging views on the best solution for the exam organization. These views have emerged from specific historical, political, and cultural traditions and beliefs. The functions the different examinations are used for range between certifying graduation, monitoring and controlling the system as a whole or individual schools in particular, and supporting schools by providing them with information. In how far the features of different examination procedures may guide and support, or alternatively undermine and forestall school development processes is a dimension that has so far been neglected in research. Given that, it appears to be paradox to speak of *the* impact of statewide exit examinations, or to assume that they are suitable per se for raising school quality and are thus the silver bullet of quality assurance. Instead, it must be discussed whether some procedures are, within their context, better suited than others to fulfill what they are supposed to, and whether in that view, the procedures maintained by some states are flawed. Answering these questions goes beyond the results of this study and shows a desideratum for research.

On that account, further research into the effects of statewide examinations on school quality will have to compare different types of examinations as to their structures and impact. This means that within the specific parameters of the education system, the intentions and expectations underlying the design of the examinations must be investigated and – in a multi-level approach – be related to actually observed effects on the actions of actors involved in constituting the product school.

[6] Here, we cannot use the example of Ireland, as the Irish education and examination system are not geared towards marketization and competition.

Acknowledgements

The presented study was conducted in the context of the Research Group & Graduate School *Teaching and Learning of Science* at the University of Duisburg-Essen. The authors would like to thank the German Research Foundation (DFG) for their support.

References

van Ackeren, I. (2003). *Evaluation, Rückmeldung und Schulentwicklung: Erfahrungen mit zentralen Tests, Prüfungen und Inspektionen in England, Frankreich und den Niederlanden.* [Evaluation, feedback and school development. Experiences with statewide tests, examinations, and inspections in England, France, and the Netherlands]. Münster: Waxmann.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives 10(18) Retrieved from: http://faculty.mdc.edu/jmcnair/arieal/Articles/High-Stakes%20Testing.htm.*

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

Bellmann, J., & Weiß, M. (2009). Risiken und Nebenwirkungen Neuer Steurung im Schulsystem: Theoretische Konzeptualisierung und Erklärungsmodelle [Risks and side-effects of new control strategies in the educational system. Theoretical conceptualization and explanatory models]. *Zeitschrift für Pädagogik, 55*(2), 286–308.

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

Center on Education Policy (CEP). (2008). *State high school exit exams: A move toward end-of-course exams*. Washington, DC.

Cole, A., & John, P. (2001). Governing education in England and France. *Public Policy and Administration, 16*(4), 106–125.

DeBray, E., Parson, G., & Woodworth, K. (2001). Patterns of response in four high schools under state accountability policies in Vermont and New York. *Yearbook of the National Society for the Study of Education, 100*(2), 170–192.

Eckstein, M. A., & Noah, H. J. (1993). *Secondary school examinations: International perspectives on policies and practice*. New Haven: Yale University Press.

Fuchs, T., & Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics, 32*(2–3), 433–464, doi:10.1007/s00181-006-0087-0.

Gillborn, D., & Youdell, D. (2000). *Rationing education: Policy, practice, reform and equity*. Buckingham: Open University Press.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Santa Monica: RAND.

Hanushek, E. A., & Wößmann, L. (2007). *The role of education quality in economic growth* (World Bank Policy Research Working Paper No. 4122). Washington, DC. Retrieved from: http://ssrn.com/abstract=960379.

Jürges, H., Schneider, K., & Büchel, F. (2003). *The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany*. CESifo Working Paper No. 939.

Keeves, J. P. (1994). *National examinations: Design, procedures and reporting*. Paris: UNESCO: International Institute for Educational Planning.

Kühn, S. M., van Ackeren, I., Block, R., & Klein, E. D. (2009). Zentrale Abiturprüfungen: Zur Heterogenität der Prüfungsverfahren in Deutschland [State-wide Abitur examinations. On the heterogeneity of exam procedures in Germany]. *SchulVerwaltung, Ausgabe Hessen und Rheinland-Pfalz, 14*(10), 185–281.

Laukkanen, R. (2008). Finnish strategy for high-level education for all. In N. C. Soguel & P. Jaccard (Eds.), *Governance and performance of education systems* (pp. 305–324). Dordrecht: Springer.

Louis, K. S., & Versloot, B. (1996). High standards and cultural diversity: Cautionary tales of comparative research: A comment on "Benchmarking Education Standards" by Lauren B. Resnick, Katherine J. Noaln, and Daniel P. Resnick. *Educational Evaluation and Policy Analysis, 18*(3), 253–261.

Maag Merki, K. (2008). Die Einführung des Zentralabiturs in Bremen: Eine Fallanalyse [The introduction of state-wide Abitur examinations in Bremen: A case study]. *Die Deutsche Schule, 100*(3), 357–368.

Maag Merki, K., Klieme, E., & Holmeier, M. (2008). Unterrichtsgestaltung unter den Bedingungen zentraler Abiturprüfungen: Differenzielle Analysen auf Schulebene mittels Latent Class Analysis [Classroom instruction under the conditions of state-wide Abitur examinations. Differential analysis on school level via latent class analysis]. *Zeitschrift für Pädagogik, 54*(6), 791–808.

Maag Merki, K., & Holmeier, M. (2008). Die Implementation zentraler Abschlussprüfungen: Erste Ergebnisse zu den Effekten der Einführung auf das schulische Handeln der Lehrpersonen [The implementation of state-wide Abitur examinations. First results on how they affect the actions of teachers]. In E.-M. Lankes (Ed.), *Pädagogische Professionalität als Gegenstand empirischer Forschung* (pp. 233–244). Münster: Waxmann.

Maag Merki, K., Holmeier, M., Jäger, D. J., & Oerke, B. (2010). Die Effekte der Einführung zentraler Abiturprüfungen auf die Unterrichtsgestaltung in Leistungskursen in der gymnasialen Oberstufe [The effects of implementing state-wide Abitur examinations on learning and teaching in upper secondary education]. *Unterrichtswissenschaft, 2*, 173–192.

National Institute for Public Education. (2003). *The system of content regulation in Hungary: Public policy analysis*. Retrieved from: ftp://ftp.oki.hu/english/Content_Regulation.pdf.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Noah, H. J., & Eckstein, M. A. (1992). The two faces of examinations: A comparative and international perspective. In H. J. Noah & M. A. Eckstein (Eds.), *Examinations. Comparative and international studies* (pp. 147–170). Oxford: Pergamon Press. Retrieved from: http://www.hku.hk/cerc/3e.html. Reprinted by permission of Butterworth-Heinemann Ltd..

State Examinations Commission (SEC). (n.d.). *Annual report 2006*. Retrieved from: http://www.examinations.ie/about/SEC_Annual_Report_06.pdf.

Vogler, K. E. (2006). Impact of an exit examination on English teachers' instructional practices. *Essays in Education 16(Spring). Retrieved from: http://www.usca.edu/essays/vol162006/vogler.pdf.*

Vogler, K. E., & Carnes, G. N. (2009, April). *Comparing the impact of a high school exit examination on science teachers' instructional practice*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Volante, L. (2007). Evaluating test-based accountability systems: An international perspective. In *Proceedings of the annual meeting of the association for educational assessment*. Stockholm, Sweden: AEA Europe.

Wößmann, L. (2003). Central exams as the "Currency" of school systems: International evidence on the complementarity of school autonomy and central exams. *CESifo DICE Report,* 4: 46–56.

Esther Dominique Klein did the teacher education program at the University of Duisburg-Essen. She has been working at the research group and graduate school *Teaching and Learning of Science* and the Faculty of Educational Sciences at the University of Duisburg-Essen, Germany, since November 2009. Her research interests lie in the area of education policy, school development research and international comparative education.

Dr. Isabell van Ackeren is a professor of educational sciences at the Faculty of Educational Sciences, University of Duisburg-Essen (Germany). Her main focus is on school effectiveness and school improvement research, education systems and policies and comparative education.