education policy analysis archives

A peer-reviewed, independent, open access, multilingual journal

epaa aape

Arizona State University

Volume 20 Number 8 March 20, 2012 ISSN 1068-2341

The Impact of Statewide Exit Exams: A Descriptive Case Study of Three German States with Differing Low Stakes Exam Regimes

Isabell van Ackeren
Rainer Block
E. Dominique Klein
Svenja M. Kühn

University of Duisburg-Essen, Faculty of Educational Sciences
Germany

Citation: Ackeren, I. van, Block, R., Klein, E.D. & Kühn, S.M. (2012) The Impact of State-Wide Exit Exams in Germany: A Descriptive Case Study *Education Policy Analysis Archives, 20* (8). Retrieved [date], from http://epaa.asu.edu/ojs/article/view/1011

Abstract: In this article we present results from a study investigating the impact of three state exit exam systems on teaching and learning in college-preparatory schools. The study compares one state with a traditionally more centralized exam regime, one state that is more de-centralized and one state that has recently switched to more centralized testing. The German Abitur is a cognitively rather complex exam that is largely unstandardized as measured by the standards of international testing regimes. Moreover, performance differences in system monitoring tests between states with different exam regimes can only be found for mathematical literacy. Therefore, centrally regulated

Manuscript received: 09/20/2011
Revisions received: 01/31/2012
Accepted: 02/16/2012

topics and grading criteria –as opposed to exams that are locally designed and reviewed centrally– make little difference in softer subjects with a more open canon, but seem to have a stronger impact in mathematics. Against this background and taking an international perspective, we argue that an overall low stakes testing regime might be the first step towards a good compromise between local flexibility for students' interests on one hand, and rigor as well as a healthy dose of performance motivation on the other hand.
Keywords: statewide exit exams; secondary school leaving certificate; German Abitur exams; low stakes testing

El impacto de los exámenes de egreso a nivel estatal: Un estudio de caso descriptivo de tres diferentes estados alemanes

Resumen: Este trabajo presenta los resultados de un estudio que investigó el impacto de los sistemas de examen de salida de tres estados en la enseñanza y aprendizaje en las escuelas que preparan para la universidad. Este trabajo compara un estado con un sistema de examen tradicional más centralizado, un estado más descentralizado y un estado que recientemente cambio a pruebas más centralizada. El *Abitur* alemán es una prueba cognitiva muy compleja, en gran medida no esta estandarizada de acuerdo con las normas de los exámenes internacionales. Por otra parte, las diferencias de rendimiento en las pruebas de evaluación de estados con diferentes sistemas fueron significativas sólo en el área de alfabetización matemática. Por lo tanto, las preguntas y los criterios de nota regulados de manera central, a diferencia de las pruebas que son diseñadas localmente y controladas de forma centralizada, tienen poca influencia en las áreas curriculares con "cánones" múltiples, pero parecen tener un impacto ligeramente mayor en matemáticas. En este contexto, y tomando una perspectiva más internacional, se concluye que un sistema general de pruebas sin consecuencias severas puede ser el primer paso para mantener flexibilidad local que sirva a los intereses de los alumnos, rigor académico y una buena dosis de motivación.
Palabras clave: los exámenes estatales de graduación, el diploma de secundaria; exámenes Abitur alemán, exámenes con consecuencias menores

O impacto de exames de saída a nível estadual: Um estudo de caso descritivo de três diferentes estados alemães

Resumo: Este artigo apresenta os resultados de um estudo que investigou o impacto dos sistemas de exame de saída de três estados sobre o ensino e a aprendizagem nas escolas que preparam para a universidade. O estudo compara um estado com um sistema de exame tradicionalmente mais centralizado, um estado que é mais descentralizado e um estado que mudou recentemente para testes mais centralizados. O *Abitur* alemão é um exame cognitivo bastante complexo que não é, em grande parte, padronizado de acordo com os padrões dos exames internacionais. Além disso, as diferenças de desempenho na avaliação dos exames estaduais com regimes diferentes só foram significativas na área de alfabetização matemática. Portanto, as perguntas e os critérios de nota regulados centralmente, ao contrário dos testes que são localmente concebidos e monitorados centralmente, fazem pouca diferença em áreas curriculares com múltiplos "cânones", mas parecem ter um impacto ligeiramente mais significativo em matemática. Neste contexto, e tendo uma perspectiva mais internacional, concluímos que um sistema geral de testes sem consequências muito severas pode ser o primeiro passo para manter a flexibilidade local que serve aos interesses dos alunos, o rigor acadêmico e uma boa dose de motivação.
Palavras-chave: graduação estado exames diploma do ensino médio; exames Abitur alemão, exames com consequências menores

Introduction

External assessment of student achievement that is based on predefined standards and centrally administered tasks plays an important role in the (mostly political) debate on instruments that can help in raising the quality of education systems. In this context graduation procedures involving statewide exit exams, especially at the end of upper secondary education, have gained more and more popularity all over the world during the past decades (cf. Klein & van Ackeren, in press). They summarize the achievement of individual pupils at the end of this educational stage, their results are used to award certificates (e.g. European Commission, 2011), and in this, the exams have a more or less significant impact on students' further careers in education, training and professions.

This movement towards statewide exit exams has been supported by changes in the governance of education systems, which become apparent mainly in increased school autonomy and more decentralized decision powers. Statewide exams at the end of the school career are seen as an instrument with which the school system can be "navigated". This view probably serves as the foundation for recent moves in various countries to either adopt statewide exit exams, as is the case for instance in Austria and the Czech Republic, or reform existing statewide exam systems with the explicit aim of assuring and improving the quality of schooling, as can be seen for instance in Hungary (see, e.g., National Institute for Public Education, 2003). Table 1 depicts the different types of exit exams for the OECD-countries. In the US, the Center on Education Policy regularly collects and reports data on state policies that require students to pass a state assessment in order to receive a high school diploma (Center on Education Policy (CEP), 2011). The authors report that twenty-five states have adopted or are preparing policies that require students to pass an exit exam to graduate from high school; many of these states participate in state consortia to develop common assessments that are aligned with the Common Core State Standards.

Table 1.
Overview of Graduation Procedures at the End of Upper Secondary Schooling in the OECD Member States

Statewide exit exams only	Statewide and school-based exit exams	School-based exit exams	No exit exams	Federal states with diverse procedures
Denmark	Italy	Iceland	Belgium***	Australia
Finland	Netherlands	Austria**	Japan****	Canada
France	Norway	Switzerland	Korea****	Germany
Greece	Slowakia	Czech Republic**	Sweden	USA
Ireland	Hungary		Spain	
Luxembourg			Turkey	
Poland*				
Portugal				
Scotland				
UK				

* Statewide exit exams compulsory only for university entrance.
** Statewide exit exams are being prepared.
*** Schools can decide to hold exit exams.
**** Statewide exams after each term and statewide entrance exams for each stage, but no specific statewide exit exams for graduation.

Normative Discourse: Intended and Unintended Impact

External assessments are regarded as being capable of ensuring that a comparable and consistent quality standard is maintained in schools (output dimension, *comparability of performance standards*) on one hand, and of moving educational processes in schools and classrooms in a desired direction (process dimension) on the other hand. This is supposed to happen in different ways (e.g. Bishop & Wößmann, 2004; Kühn, 2010; Klein & van Ackeren, in press): Exams that are administered by the state are expected to ensure that teachers cover the full scope of the curriculum, and even contents that they are unfamiliar with, as they do not know in advance the contents and foci of the exam tasks *(coverage of content standards)*. Statewide exams are also believed to have the potential for promoting the implementation of new and modern curricula and task formats (e.g. accounting for authentic contexts and student's life experience, problem solving, cross-curricular competences etc.) *(innovation potential)*. In this respect, it is also expected that the tasks (as regards content and format) that teachers use in the classroom approximate the quality of exam tasks. Taking into account that teachers have no access to actual exam items and tasks, this kind of 'alignment' might, from an optimistic perspective, help to ensure some basic quality standards of modern instruction, e.g. regarding "real world" orientation *(alignment of testing and instruction)*. A key precondition for this kind of meaningful orientation towards the tests (instead of negative "test coaching" or "teaching to the test" effects) are high quality tasks that do not simply ask to recall facts and information (Longo, 2010). In addition, the exams are said to enhance the commitment of both students and teachers by increasing their extrinsic motivation as they mutually face a challenging situation *(extrinsic motivation)*. The exams are also supposed to promote the diagnostic skills of teachers and their use of criterion-based assessment *(criterion-referenced rigor)*. The rationale behind this is that students should be graded according to the actual quality of their work rather than by reference to how other students in a course perform on the same tasks (Sadler, 2005). Finally, it is believed that both the students themselves and the institutions and organizations that recruit graduates have a better instrument to gauge the individual abilities of graduates when final grades are based on a statewide exit exam instead of school-based exams or assessments *(public acceptance)*.

However, despite the popularity of statewide exit exams and the positive effects that are usually ascribed to their introduction, we must assume that these exams may also have unintended side effects (van Ackeren, 2007), in that the delivered curriculum is narrowed down to announced test contents and formats ("teaching to the test" as a negative form of *alignment of instruction and testing*); that the use of reproductive learning is increased while comprehension-oriented and creative forms of learning are marginalized –an aspect that is probably associated with the quality of the tasks–; that current topics, local conditions, and teacher/student interests are considered less *(flexibility, individual student or local norm reference)*; and that teachers feel de-professionalized since some of the status of teachers as experts for learning, instruction and assessment with a specialized knowledge and a high standard of professional ethics, behavior and work activities, is cut *(teachers' professional self-concept)*.

In Focus: German *Abitur* and Differing Exam Procedures

To illustrate the "turn towards statewide exams", the situation in Germany can serve as an example. In Germany, many federal states have switched from school-based to statewide Abitur exams after 2005. Before 2005, only about half of the German states conducted statewide Abitur exams –partly because the occupying allied powers after 1945 installed the exam systems that were common in their countries, and partly because the former GDR states carried on with their comprehensive school tradition (Polytechnische Oberschule) after the reunification–; in the other half, the Abitur was granted on grounds of exams that were set by course teachers and accredited by

the respective ministry of education. Following debates about the quality of the education system that were particularly fuelled by the German results in PISA and other large-scale assessments, all states of the latter group—with the exception of one state, Rhineland-Palatinate[1]—have replaced their school-based with statewide Abitur exams (Klein et al., 2009).

The Abitur was established in 1834 as a graduation exam for the upper level of the Gymnasium that granted admission to university. It has since grown beyond that purpose, and increasingly become a prerequisite to start an apprenticeship in some professions as well. The academic level of the Abitur is comparable to the International Baccalaureate, which provides an internationally accepted qualification for entry into higher education. The Abitur is the only school-leaving certificate in all German states that allows students to go on to university directly; however, some universities set their own additional entrance exams. Equivalent graduation certificates in other countries are the Matura (e.g. Austria) or A-levels (e.g. England, Wales, Northern Ireland, Hong Kong, Singapore).

To be admitted to the exam, certain requirements have to be met during the qualification phase: From the tray of subjects students are enrolled in, they have to choose four or five exam subjects, which must include two of the following three subjects: German, a foreign language, and mathematics. The exam subjects are taught at different levels of academic standards in accordance with the Uniform Examination Standards in the Abitur Examination (Einheitliche Prüfungsanforderungen in der Abiturprüfung – EPA); two of the four exam subjects must be studied at a level of increased academic standards. The further details of the exam process are in the responsibility of each state.

The instruction in the upper secondary level is supposed to provide an in-depth general education, foster the general capacity for academic study, and prepare students for scientific work (European Commission, 2011). Therefore the exit exams tend to include cognitively complex tasks that require competences beyond knowledge recall. One specific characteristic of the German exams is that in both types –statewide and school-based– the papers are marked and scored by local teachers according to pre-specified grading criteria. Two teachers mark the papers independently, and in this have to interpret the grading criteria, judge a complex piece of student work, and assign the grade. This is starkly different from most testing regimes commonly discussed in the literature, and especially in the American literature (see below, "state of research").

Objectives of the Study

In Germany as well as other countries in Europe, the graduation systems have been changed and modified with the explicit or implicit understanding that the exams can contribute to the improvement of school quality. On one hand, the exams are supposed to 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 strategy)*. On the other hand, they hold school members accountable for their work, and link the graduation of the students to the exams *(pressure strategy)*. Especially regarding the latter, it should be noted that in contrast to the well-established accountability systems in the US, the European systems are mostly conceptualized as rather low stakes systems including only some high stakes elements. Altogether, empirical results

[1] The governance philosophy of Rhineland-Palatinate focuses on a rather supportive understanding of quality improvement for schools. Furthermore, based on political consulting, the ministry of education justifies its decision to abstain from statewide Abitur examinations with a lack of consistent empirical evidence about the intended and unintended effects of these exams (documented in diverse press releases and articles). Finally, Rhineland-Palatinate has had very positive results in PISA in the last years; therefore, public pressure has been rather absent.

that can either support or confound the assumptions about how the new systems can affect schools, especially with regard to the process dimension of school and classroom work, are pending.

In this context, the objective of the study described here is to analyze how educational processes in upper secondary schools are affected by statewide vs. school-based exit exams in an overall low stakes school system. This is done in that three German states with different exam organizations are juxtaposed and the effects are compared through quantitative questionnaire data. In the following, we first outline the current state of research on the effects of exit exams with different organizations. We then describe the design of our study, and give a detailed description of the statistical evaluation methods that we are using. After that we summarize our findings, and conclude with a critical discussion of the results and their practical relevance for education policy.

State of Research

Empirical findings on how statewide exams affect student achievement and instruction are both sparse and inconsistent. In the following illustration of existing research on the impact of statewide exit exams, we distinguish between studies that focus on learning outcomes on one hand and the impact on instruction processes on the other hand.

Statewide Assessments and Learning Outcomes

Various German and international studies investigate student performance on standards-based (exit) exams and tests with research approaches that are rooted in educational economics. Some of these studies suggest a causal relationship between statewide exams and student outcomes (cf. e.g. Bishop, 1998; Cosentino de Cohen, 2010; Jürges, Schneider, & Büchel, 2003; Fuchs & Wößmann, 2007; Hanushek & Wößmann, 2007; for a summary, see Wößmann, 2008). On the whole, this research indicates that statewide exit exams tend to have positive effects, but the picture that emerges across different studies is mixed at best –the findings are to some extent contradictory and vary across subjects, course levels, and age groups (Maag Merki, 2008). The German results in TIMSS (data collected in 1996, Baumert & Watermann, 2000) for instance, indicate that the statewide Abitur –compared to the school-based Abitur– seems to be able to ensure standards in mathematics achievements in the lower performance ranges. A comparable result, however, cannot be found for physics, the other subject that was tested. The authors assume that the exams can only unfold their potential of assuring minimum standards in compulsory courses such as mathematics, which at lower performance level cannot be dropped by students in Germany, whereas it is possible to drop physics in general and advanced level mathematics so that these courses are probably chosen by rather higher performing and more motivated students.

Using data from TIMSS and PISA, Wößmann (2003, 2008) analyzes whether and how student achievements in different countries are connected with the according education system and graduation procedure. Even though Wößmann's results seem to confirm the potential of statewide exams for ensuring standards on a statistical basis, it should be noted that he uses achievement data of 13- and 15-year-olds. This age group is not directly affected by the exams, which take place at the end of upper secondary schooling, when students usually are about 18 to 19 years old. Therefore we can only—if at all—suggest a "distant effect" of the exams on their achievements (cf. Schümer & Weiß, 2008). Our own reanalysis of the German PISA-E 2003 study[2] shows that when the

[2] PISA-E was an expansion of the PISA-tests in 2000, 2003 and 2006 and tested about the tenfold amount of students of the regular PISA study in Germany; in 2003, only seven of the sixteen German states had statewide Abitur examinations.

achievements of German states with statewide exit exams are contrasted in pairwise comparisons with those of states without this type of exam, we can find neither a statistical superiority nor a practically relevant general supremacy of states with statewide exams (cf. Block et al., 2011). However, from a subject-specific view we find statistically significant performance differences between students of statewide and school-based exam regimes in mathematics, but not in reading and science literacy. When the mathematics performance of students is adjusted for meaningful background variables (sex, ethnicity, socio-economic status), we find that three of the seven states with a statewide exam regime (Bavaria, Saxony and Baden-Württemberg) are higher performing with a small to medium effect than three to five states without statewide exam procedures. Rhineland-Palatinate, which is in focus in the study we present here, is not part of the lower-performing states without statewide exams.

The case study we present in this text analyzes the impact of statewide exams on school and classroom practice –as the link between exam regime and student performance– in selected states with differing exam regimes. The performance of these particular states can be illustrated with PISA-E 2006-data (see table 2). There are no substantial differences between the traditionally more centralized system in Baden-Württemberg and the more decentralized exam tradition in Rhineland-Palatinate except in mathematical literacy. This seems to be in line with the subject-specific results of the other studies presented above.

Table 2.
Mean Student Performance in Baden-Württemberg and Rhineland-Palatinate, PISA 2006

States and Reference Groups	Science Literacy		Reading Literacy		Mathematical Literacy	
	M	SE	M	SE	M	SE
Baden-Württemberg	523	(2.8)	500	(4.2)	516	(3.2)
Rhineland-Palatinate	516	(2.8)	499	(3.0)	499	(3.0)
Germany	516	(3.8)	495	(4.4)	504	(3.9)
OECD	500	(0.5)	492	(0.5)	498	(0.5)

Data retrieved from PISA-Konsortium Deutschland, 2008

Statewide Assessments and School / Classroom Practice

Studies looking at the impact of statewide exams on instructional processes –and thus on how teachers and students, who are directly involved in instructional practice, respond to the exams– are limited, and the findings at hand are largely inconclusive (see Maag Merki, 2008; Maag Merki, Klieme, E., & Holmeier, 2008). Current results from the two German states Bremen and Hesse show that the statewide Abitur exams positively influence some dimensions of instructional quality in advanced level mathematics and English courses, but not in German and biology (Maag Merki, Holmeier, Jäger, & Oerke, 2010); the comparison of results for each of the two states indicates that the quality of instruction has improved in Bremen after statewide exams have been introduced, whereas there seems to be a slight decline in instructional quality in Hesse. The authors conclude that further studies need to test how stable these outcomes are.

The innovation potential of test and exam tasks has been examined by Kühn (2010; in press) for the science subjects –biology, chemistry, and physics– in the exit exams of selected German states and other countries. Especially in science education, many innovation-oriented concepts have recently been discussed for improving instruction and the corresponding tests and exit exams. However, in the context of implementation research, Kühn states that in most aspects of the exam tasks explored in her study, there is only little coherence between what the externally determined requirements for the construction of exit exam tasks call for and how this is actually carried out in the task practice. Accordingly, the opportunity to implement newer concepts in the exam task design actually seems to remain largely unused.

Most studies from the US focus on standardized tests (see, e.g., Au, 2007; Volante, 2007) which –in contrast to German tests– usually are part of an elaborate school accountability system and therefore have high stakes attached for the schools and individual teachers as well as for the students involved, whereas German exit exams have high stakes for the students only. The majority of these studies have found that tests with high stakes for the schools often have narrowing effects in terms of both the delivered curriculum and the teaching methods used: Attention is focused on the subjects tested, and there is little capacity within these subjects to respond to students' interests. In the direct run-up to the tests, in particular, there is an increased focus on task formats similar to those used in the tests, and class discussion centers on contents typically addressed in the tests (e.g., Firestone, Mayrowetz, & Fairman, 1998). However, the results can only partially be transferred to the situation in Germany. It is conceivable that the distortions frequently found in US studies are associated with the characteristics of the testing regimes investigated, which often are highly standardized, oriented towards basic skills, challenging especially for low-achieving students, externally scored, connected to high pressure, and possibly sometimes not well aligned to curricular content. Only a few studies from the United States focus high school exit exams in particular, and most of them investigate student achievements and drop-out rates (for an overview, see Holme 2010). Some qualitative studies zoom in on how district and school work is influenced by the exams, and indicate that how schools and districts respond to the exams depends on local capacities and aspects of school or department culture to a large degree (cf. e.g. Center on Education Policy (CEP), 2007; DeBray, 2005; Goertz & Massell, 2005).

To summarize, previous studies could not identify any common effects of statewide exit exams on process and outcome variables. Indeed, several review studies of how exams are organized in different states (for a summary, see Klein et al., 2009) show that the seemingly generic label 'statewide exam' is applied to rather diverse approaches. A comparison of Abitur exam procedures in Germany reveals comparatively few elements that are common to all states. Moreover, compared with procedures in other European countries, the procedures in most German states are only moderately standardized and have a low stakes character. Against this background, we do not expect the distortions that can be found in the US to also show up in the German states.

Besides this, most of the studies described above draw on a one-dimensional model of causes and effects, and thus fail to recognize that student achievement is influenced by a multitude of factors and conditions. They often also neglect how the education systems that the exams are embedded in are organized (for criticism of this approach, see Block et al., 2011; Maag Merki, 2008; Schümer & Weiß, 2008).

Study Design and Methods

The aim of the research project "Conditions and Effects of School-Based and Statewide Exit Exams" is to explore the diverse influences of different forms of exit exams on educational

processes and outcomes in Germany; three smaller studies have been carried out as part of the overall main study (a systematic survey of German and international exam procedures, the re-analysis of PISA performance data described above, and a comparative study of tasks in statewide and school-based exams). The project scrutinizes several assumptions that have been phrased in the literature about how the levels of school system, individual schools, instruction, and individual stakeholders influence the impact of statewide exit exams on student achievement, but remain largely unsubstantiated. In this, the project focuses exam regimes that can be characterized as low stakes, especially since the grading of exit exams is done locally in all German states, regardless of whether the exams are statewide or school-based. From an international perspective this might be an interesting point given the many unintended effects of high stakes accountability systems that have been found in mainly American studies.

From the largely normative assumptions in the literature (see above), we have extracted a few key dimensions and a list of indicators for intended and unintended effects of statewide exit exams (see Table 3). The indicators have been further elaborated through interviews with experts in the state ministries of education, and fine-tuned after piloting in autumn 2008.[3]

Table 3.
Conceptual Model: Key dimensions

Process (instructional and individual level)	Outcome
Innovation potential*	Comparability of performance standards**
Coverage of content standards	Public acceptance***
Alignment of testing and instruction	
Flexibility (individual student / local norm reference)	
Rigor (criterion reference)	
Pressure to perform according to this rigor	
Extrinsic motivation	
Teachers' professional self-concept	

* cf. Kühn, 2010 and "state of research" in this article
** cf. Block et al., 2011 and "state of research" in this article
*** This outcome dimension will be part of future inquiry.

The article at hand investigates whether exams labeled as "statewide exit exams" under the specific German conditions affect schooling in the way they are expected to (see above). The effects are contrasted with the effects of the school-based exit exams. In this context, the following research questions are considered: How do different forms of exit exams in rather low stakes exam systems influence (1) teaching and learning processes inside and outside school (concerning breadth of the

[3] Note: Henceforth, the term construct describes the superordinate theoretical category of particular variables. Instead of scales, the report of the results contains manifest variables that conform to the selection criteria described below (see paragraph on analysis procedure).

delivered curriculum, pace of instruction, individualized instruction, consideration of students' interests, elaboration strategies, intensity and type of exam preparation, grading, homework, private tuition, parental influence), and how do they influence (2) teachers' perceived professional self-concept and students' motivation to learn (concerning type of motivation and role perception)? How do potential effects differ (3) across subjects, (4) across selected German states with different exam traditions and (5) within a state that is changing from rather school-based to rather statewide exams?

We assume that statewide exams with centrally administered exam tasks (using the examples of Baden-Württemberg and North Rhine-Westphalia), in contrast to school-based exams with locally designed tasks (by the example of Rhineland-Palatinate), prompt diverse changes in instruction and extracurricular activities. Although there is no difference between the states concerning the grading, which in both cases is done locally by teachers, we expect substantial differences. We base our assumption on the perceived relevance and general acceptance of a so-called "external" exam, which probably conveys the impression of higher standards, increased comparability and objectiveness, and independence from the teachers' notice of the test items and tasks. From this perspective, school actors might feel a higher pressure to prepare for the exams more systematically. We therefore expect the preparation for the statewide exams to be characterized by a narrowed delivered curriculum, a faster pace of instruction, a decreased capacity to respond to students' interests, less individualized instruction, a marginalization of comprehension-oriented learning, an overuse of reproductive learning, more exam preparation outside the classroom, the use of more criterion-oriented grading and less consideration of individual development or social references, an intensified homework practice, more private tuition and a stronger parental influence, an increased feeling of de-professionalization (teachers) and less intrinsic motivation (students in particular). In line with existing research, we also expect subject-specific differences as well as deviations between the selected German states (for further explanation see chapter on "Causal Attributions in Cross-Sectional Designs").

Sample and Data Collection

The descriptive case study reported in this article is based on quantitative empirical research. A cross-sectional analysis with retrospective elements has been carried out in three German states, each with differing Abitur traditions: Baden-Württemberg (BW), where the Abitur exam has been statewide since 1946, Rhineland-Palatinate (RP), where the exam is still school-based, and North Rhine-Westphalia (NW), which has introduced statewide Abitur exams in 2007. The sample comprises four schools of the academic-track[4] per state. In NW, four comprehensive schools are included in the analyses to explore potential differences between school types.[5] The participating schools are deliberately selected from socio-structurally homogeneous catchment areas in urban centers. The study is based on a full-sample survey of the teaching staff (including the school management) and students of the Abitur class of 2008. Standardized questionnaires are used to assess respondents' perceptions and evaluations with regard to the research questions as outlined above.

[4] All German states have tracked school systems with two or more different tracks in secondary education. The academic track, Gymnasium, which exists in all 16 states, and the comprehensive schools, which only exist in some states, are those school types in which students can take the Abitur and gain entrance to university.

[5] For lack of space, the results of the school type comparison are not displayed.

The survey was conducted in late 2008/early 2009, before the students took their Abitur exams. The overall response rate was 67% for students and 31% for teachers.[6] A total of 250 teachers with teaching experience at upper secondary level (54 in BW, 96 in RP, and 100 in NW, of whom 63 were in academic track schools) have been recruited. The adjusted student sample used in the case study comprises 887 students (251 from BW, 249 from RP, and 387 from NW, of whom 235 were in the academic track). As a result of structural differences in the organization of exit exams in the three states, the subject-specific analyses are based on advanced-level mathematics, German, and biology courses, the latter being the science subject most frequently taken by the students in the case study. The students' exam subjects are as follows: BW (mathematics 250, German 251, biology 47), RP (mathematics 81, German 100, biology 129), NW academic-track schools (mathematics 84, German 80, biology 38), NW comprehensive schools (mathematics 69, German 66, biology 57).

Causal Attributions in Cross-Sectional Designs

We cannot draw any generalizable conclusions or causal inferences about different forms of exams from the present small-scale case study with a cross-sectional design. However, descriptive approaches are an indispensable component of the cumulative research process: "Case studies –and particularly individual case studies– serve primarily explorative purposes in the context of 'quantitative' social research: a specific area of social reality … is to be defined in descriptive terms" (Kromrey, 1998, p. 507, our translation). We therefore took several steps to ensure that we can draw plausible conclusions from our findings. To minimize the effects of unobserved heterogeneity, we deliberately selected the school sample from socio-structurally homogeneous catchment areas in urban centers. Furthermore, homogeneity tests did not reveal any statistically significant differences in the participating teachers across the states under examination.

Moreover, the three states were deliberately chosen to maximize the variance of the observed data. BW not only has one of the longest traditions of statewide exams among the 16 German states, but the exam procedures implemented show a relatively high level of standardization in the German context (e.g., statewide exams in all written Abitur subjects; limited choice for students and teachers; use of a second, external examiner; at least partial anonymity of exam candidates in the grading process; see Klein et al., 2009). RP, with its school-based exam procedures, does not have any experience with statewide exams at all. NW has recently changed from one exam type to the other; therefore, the rating of our indicators in this state can be expected to fall between those of the other two states. Regarding the different traditions within these states –and provided that statewide exams do actually have an effect– , if ratings are higher or lower in NW than in BW and RP, this might point to either overcompensation or inertia effects during the transition to statewide exams in NW. Moreover, findings may indicate that how NW teachers' retrospectively evaluate changes that have come up during the transition confirms and substantiates the differences identified between BW and RP. This would strengthen the plausibility of a causal association with the organization of exit exams and its perceived importance and demand, respectively. Given the

[6] There are huge differences between the response rates of individual schools. The response rates for student questionnaires vary between 60 and 90% with two schools having lower rates. Regarding teacher questionnaires, the response rates range from 20 to 50% per school, with two lower outliers as well. Due to the intensified quantitative empirical research on school effectiveness during the past decade, and in the light of the implementation of diverse test and inspection systems for German schools, many teachers seem to be tired of additional surveys that are conducted by university research teams. Especially with regard to this small teacher sample, the results only allow for cautious interpretations.

comparatively low number of participants in the case study, moreover, we use rigorous statistical selection criteria to guard ourselves against prematurely identifying exam effects (see below).

Analysis Procedures

We choose a data mining and knowledge discovery approach (Hastie, Tibshirani, & Friedman, 2009; Witten & Frank, 2005) to select indicators that can be considered relevant in terms of educational governance and policy. Statistically relevant indicators are identified in a multi-stage procedure. Of the more than 300 indicators rated by the teacher and student sample, only those that meet the following three criteria are identified as representing potentially differential effects of statewide or school-based exit exams in BW and RP, respectively:

Significant V-test score (significance level: alpha = .05; reference group: BW).

The V-test score indicates whether an indicator was rated significantly higher or lower in the reference group than in the total sample (Lebart, Morineau, & Piron, 2000). The rule of thumb is that absolute values > 2 indicate a significant difference. The larger the absolute V-test score, the larger the difference. When two groups are compared, as in the present analysis of BW and RP, the procedure is comparable with a classic t-test. In the data mining approach, the results of probability tests have a heurostatistical, exploratory character and explicitly do not serve the purpose of testing population hypotheses (alpha levels are therefore not adjusted).

At least moderate differences in effect sizes.

An eta^2 statistic ≥ .06 is taken to indicate a moderate effect size. As a rule, eta^2 values indicating moderate effect sizes of indicators with a significant V-test score also prove to be significant. When two groups are compared, the eta^2 effect size of the standardized mean difference is equivalent to *d*. Only mean differences between BW and RP of at least a moderate effect size are reported and interpreted as practically relevant, substantial mean differences. For one thing, Monte Carlo simulations have shown (see Barnette, 2006) that –depending on the sample and group size of random samples– small effect sizes frequently occur even when there are no mean differences between groups. Given the relatively low number of participants in the case study, we seek to guard against prematurely identifying exam effects on the basis of small mean differences. For another thing, education policy and, to an even greater extent, approaches of educational economists expect statewide exams to have pronounced effects on educational practice and outcomes.

More variance is explained at the state level than at the school level.

The necessary –if not sufficient– condition for statewide exit exams to have the intended effects on educational practice is met only if more variance is explained at the state level than at the school level. Because the eta^2 statistic does not take into account the hierarchical, clustered structure of the data, we also report the intraclass correlations (ICC) for the state level (level 3, BW and RP) and the school level (level 2, the individual schools). For dichotomous target variables, we further report the Rho intracorrelation. The ICCs are estimated from multilevel random intercept only models using the REML method and indicate the amount of variance in the indicator under investigation that is explained at the state level and at the school level in percent. Likelihood ratio tests are used to compare the deviance of 3-level and 2-level random intercept only models (i.e., models that do/do not take the state level into account). The ICCs of the third level, which are of primary interest here, are shown in parentheses in the tables if the random intercept only model with three levels of analysis and the random intercept only model with only two levels of analysis (i.e., without the state level) do not differ significantly in the likelihood ratio test (at a significance level of

alpha = .05). Given the comparatively low number of participants in the case study, however, the results of the significance tests for the multilevel analyses should not be over-interpreted, and the ICCs should be viewed primarily as descriptive measures.

Only indicators meeting all three of these statistical criteria are interpreted as being potential effects of the exam organization. Finally, we report how NW teachers retrospectively evaluate changes brought about by the introduction of statewide exams only when at least 50% of respondents identify a change as having occurred and the frequencies are significantly different from a uniform distribution (using a significance level of alpha = .05).

Results

Student Survey

We begin by presenting findings on the potential influence of statewide exit exams in specific subjects, looking first at advanced-level Abitur courses in mathematics. In our analyses we focus differences between BW and RP (see above). Table 4 lists the instruction indicators by the magnitude of the V-test statistic – the higher the absolute value of the statistic, the more pronounced the differences between BW and RP. The table presents those indicators that differ significantly between BW and RP, with the difference translating into at least a moderate effect size, and in which more variance is explained at the state level than at the school level. Non-significant ICCs are reported in parentheses, and –where reported– those mean ratings for NW that are outside the range of the two 'contrast groups' BW and RP are shown in bold (see Table 4). Apart from that, deviations as well as congruencies that occur to be especially relevant or interesting will be discussed in addition to being displayed in the table.

In the schools of our sample, BW and RP differ in terms of students' –subjectively perceived– ability to cope with the demands of their Abitur courses (perceived pressure and ability to perform according to given rigor). This is reflected in reports of a faster pace of instruction and more intensive homework practice in BW. It seems that teachers maximize classroom learning time and opportunities to cover Abitur course content by setting more homework. In the core subject mathematics, the academic-track students surveyed in BW have a latent fear of falling behind in the lessons and in their homework. Accordingly, a higher proportion of students in BW report that they get private tuition in mathematics outside the school at upper secondary level.

Substantial differences are also apparent in forms of classroom learning (alignment of testing and instruction). According to the students surveyed, instruction in mathematics in BW contains less group work and social learning, fewer opportunities to discuss contents with fellow students (or with the teacher), and students having less say in the topics covered (flexibility, individual / local norm reference). The classroom activities also contain less elaboration strategies when students work on lesson contents.

It is important to remember that these descriptive findings are not necessarily associated with the organization of Abitur exams in the two states. Nevertheless, indicators that are corresponding with the hypothesized effects of the statewide Abitur (e.g., marginalization of social and creative forms of learning because they are less relevant for exam tasks; decreased capacity to respond to students' interests when announced thematic core areas have to be covered) are particularly strongly endorsed in BW relative to RP. At the same time the means for NW (the state in transition from a school-based to a statewide exam procedure) are between the means of BW and RP. Finally, it should be remembered that, for all of the indicators reported, more of the variance is explained at the state level than at the school level.

Whereas the findings for mathematics confirm several of our hypotheses, there are no substantial differences between BW and RP in advanced-level German. This does not seem to result from a lack of statistical power: The number of students in our sample taking German is larger than of those taking mathematics (see above). A similar pattern emerges for advanced-level biology. Here, too, we find almost no differential effects. One exception is homework practice, which is again reported to be more intense in BW than in RP. Thus, in the student survey, systematic descriptive differences between BW and RP can be found only for mathematics, and not for German and biology. As a first indication, we can cautiously conclude that the potential impact of statewide Abitur exams is primarily subject-specific. In a next step, we shift the focus to indicators of instructional practice that are not subject-specific (see Table 5).

It seems that students surveyed in BW are more likely than their peers in RP to prepare for exit exams outside the classroom –from working on past papers on their own to paying for private preparatory courses outside school. Moreover, students in BW report lower interest in their Abitur subjects. Relative to their peers in RP, students in BW feel that teachers tend to narrow the delivered curriculum as a means of exam preparation. In terms of exam preparation, students in NW only rate one indicator higher than their peers in BW: preparing for the exam outside the classroom by working on past exam papers. Given that NW students might probably experience uncertainties during the period of transition to the new and unknown statewide Abitur, this finding seems quite plausible.

To sum up, our analysis of student data indicates that statewide Abitur exams potentially affect processes in mathematics, but not in German or biology. For the most part, the descriptive differences detected between the states in mathematics are consistent with what was expected (e.g., a tendency to marginalize collaborative forms of learning; less room to cover current topics or to respond to local conditions and student interests). Moreover, students in BW report a comparatively high pace of instruction and more intense homework practice, and they are more likely to prepare for their Abitur exams outside the classroom.

Teacher Survey

Table 6 illustrates the general indicators that are not subject-specific and in which BW and RP differ substantially in the results of the teacher survey. In the assessment of student performance, the teachers in BW are less likely to consider individual development and social comparison in addition to a criterion-based reference than are their colleagues in RP, in both basic and advanced courses (criterion-referenced rigor). Teachers in BW reported intensive exam preparation, which is in line with the expectations; but there is no evidence that the delivered curriculum is being narrowed profoundly and persistently in BW (coverage of content standards). Parents seem to influence instruction a little stronger in BW. However, in the context of a statewide Abitur, students and teachers appear to be closer together as 'allies' facing the challenges of the exam than in the school-based Abitur, where the teacher sets the exam. Although the statewide Abitur is specifically used to ensure comparability, the existing organization does not seem to satisfy teachers in BW, as they would be very much in favor of a third examiner being involved in the exam process in addition to the two examiners who are already involved in the marking process in BW. This may be because teachers have a rather skeptical view of the rectifying role the second examiner is supposed to have. However, it may also reflect the stronger emphasis on 'absolute' objectivity in the BW teachers' approach to teaching (see Table 4). The means of teachers in NW exceed the level in BW mainly for two indicators: the intensity of exam preparation in the lessons and the perception of students' and teachers' roles as 'allies.' The increased level of exam preparation, in particular, can probably be interpreted as an implementation effect because directly after the introduction of the

statewide Abitur, teachers are still uncertain about how to behave within these changed circumstances.

Overall, according to the teachers surveyed, the potential effects of the statewide Abitur in which BW and RP differ seem –for the most part– to be functional and in line with what was intended: In particular, this means that teachers use criterion-based references rather than drawing on individual development or social comparison when grading, combined with striving for the highest possible level of objectivity in grading.

Findings for North Rhine-Westphalia

The changes identified by the majority of academic-track teachers in the schools are presented in Table 7[7]. Remembering what effects were intended by the education authorities (as elaborated in the document analysis and interviews), some of these changes can be regarded as intended and functional (e.g. concerning teacher cooperation, use of exam results and reduced workload, also see introduction chapter).

At the same time, that teachers experience increased stress and de-professionalization (in the sense that their competencies have been reduced) are changes that cannot necessarily be described as positive. By far the most frequently named change at the transition to the statewide Abitur in NW is that the delivered curriculum is narrowed (as a reminder: in the cross-state comparison, narrowing of the curriculum was an indicator discriminating between BW and RP in the perspective of neither students nor teachers) (coverage of content standards). A thematic focus is of course intended to a certain degree; in this respect, what is relevant here is to what level and how long the curriculum is narrowed. The narrowing might possibly be due to overcompensation at the transition to the statewide Abitur (see Figure 1). Ratings for this indicator are markedly higher in NW than in BW, where (with the exception of the period of three months before the exam) they were at a level that was similar to the level in RP, where the Abitur is school-based.

In this context, findings on aspects that teachers in NW feel have not changed substantially since the introduction of the statewide Abitur are just as informative for estimating the capacity of the statewide Abitur to improve the quality of schooling. In the ratings of our teachers, these aspects include homework practice (no significant change is perceived in the amount or intensity of homework or in the type of tasks set) as well as the grading and evaluation of student performance. Grading evidently is no stricter for externally set tasks than for internal tests. Neither the relevance of the second examiner, nor the grading norm (criterion, individual development, social comparison) used in the evaluation of student performance, is perceived to have changed. According to the academic-track teachers, learning is no more teacher-centered than before, and parental involvement is no higher. These findings can be interpreted as evidence of inertia or persistence effects during the implementation of the statewide Abitur.

Summary and Discussion

In this article we have presented results from a study on the impact of three state exit exam systems on teaching and learning in college-preparatory schools (with a special focus on the alignment of instruction with testing). The study compares three different exam regimes in Germany, one state with a more centralized exam, one state with a more de-centralized exam, and one state that has recently switched from more de-centralized to more centralized exams. However,

[7] The table reports percentages instead of mean and standard deviation to assure comparability between these and other variables that are not interval scaled.

from an international point of view, the centralized exams in Baden-Württemberg and recently in North Rhine-Westphalia still appear to be rather de-centralized, for instance by the standards of a typical American testing regime, especially because of the local grading procedure. Nevertheless, we assumed that the anticipation of a "central exam" (Zentralabitur) results for instance raised expectations and a perceived pressure to align instruction with contents and task formats of the exams. At the same time we did not expect the same amount of un-intended effects as has been documented especially for high-stakes testing in the American education system.

Of the instructional indicators surveyed, relatively few prove to differ between the more centralized testing regime in BW and the more decentralized system in RP. Given the many expectations bound up in the statewide Abitur exam, these findings can be cautiously interpreted as indicating that the statewide Abitur in the variant that is used in BW probably has rather limited (or rudimentary) impact on schooling. The finding that the potential effects differ markedly depending on the subject (marked effects in mathematics, negligible effects in German and biology) suggests that statewide exams have subject-specific rather than general effects. This interpretation is congruent with findings from other studies (e.g. Kühn, 2010), although possible reasons need to be examined in further research with a focus on department cultures. What is more, the ratings of students on one hand and teachers on the other hand seem to vary in some aspects (e.g. concerning the perceived narrowing of the delivered curriculum). This aspect should be pursued in further analyses of the data.

Ratings of the instructional indicators in NW –a state in transition between two exam types– fall consistently between those for BW and RP, as representatives of states where exit exams have traditionally been statewide or school-based respectively. Only a few indicators have higher ratings in NW than in BW (e.g., intensity of exam preparation; narrowing of the delivered curriculum). These findings can probably be interpreted as evidence of overcompensation during the transition to statewide exit exams.

For the most part, the differences identified between BW and RP are *not* congruent with the changes that teachers in NW identify in retrospect as having been brought about by the introduction of the statewide Abitur. It might be that the effects of the governance instrument statewide Abitur are overshadowed by overcompensation (e.g., coverage of content standards) and inertia effects (e.g., grading references) while it is still being implemented in NW. Finally, in the comparatively few domains in which the statewide Abitur does seem to have an impact on educational practice, most of the effects observed seem to be intended.

To a large extent, our results are in line with previous empirical findings on statewide Abitur exams in Germany. Evidence seems to be accumulating to let us suggest that statewide exams affect the subjects of mathematics (cf. Baumert & Watermann, 2000) (and English) –at both basic and advanced level– but not German or biology (cf. Maag Merki, 2008). Given the fact that the German Abitur exams are cognitively rather complex exams that are largely unstandardized compared for instance with the type used in Anglo-Saxon countries, and also given that large performance differences on system monitoring tests (e.g., PISA) between BW and RP can only be found for mathematical literacy, we can conclude that regulating topics and grading criteria centrally instead of locally with only a central review makes little difference in softer subjects with a more open canon, but seem to have a stronger impact in Mathematics. The discussion of possible subject- and level-specific effects of statewide exams is still in its infancy, however; further research is clearly needed.

To conclude, doubts remain as to the general capacity of the statewide Abitur to bring about the changes intended by the education authorities – especially in view of the diverse forms in which exit exams are currently used across the German states. This limited impact may be attributable to the rather low overall level of standardization of exam procedures in Germany in international

comparison. However, it may also be attributable to the fact that in Germany, statewide exams hold high stakes for students, but not for schools and teachers. Yet precisely this evidently spares the system the considerable 'collateral damage' of high stakes for schools that can be observed in other countries (Nichols & Berliner, 2007). Thus, an overall low stakes testing regime might signify a compromise between local flexibility that can respond to students' interests and needs, as well as rigor, and a sound amount of performance motivation. Regarding the comparability of performances between Baden-Württemberg an Rhineland-Palatinate one could reason that the exam conditions found in RP seem to be more advantageous because teachers seem to be more flexible in their teaching practice, and students are engaged and do not need to get private tuition outside the school to keep up with the demands and pressure. On the other hand we can detect positive effects in the more centralized system in BW especially when it comes to assuring minimum standards in specific subject domains. Finally, we should not underestimate the public acceptance of so-called "central" exams, especially from the perspective of future employees of the graduates. Especially because of this last point, it seems very unlikely that the federal ministries turn back the clocks and return to more decentralized exam systems.

References

Ackeren, I. van (2007): Zentrale Abschlussprüfungen. Entstehung, Struktur und Steuerungsperspektiven [Statewide exit exams: Origins, structure, and perspectives for educational governance]. *Pädagogik, 59*(3), 12-15.

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

Barnette, J. J. (2006). *Effect size and measures of association.* Retrieved January 25, 2012 from http://www.eval.org/SummerInstitute/06SIHandouts/SI06.Barnette.TR2.Online.pdf

Baumert, J., & Watermann, R. (2000). Standardsicherung durch die Abiturprüfung: Zentralabitur oder dezentrale Prüfungsorganisation? [Assuring standards through the Abitur exam: Statewide Abitur or school-based organization of examinations?] In J. Baumert, W. Bos, & R. Lehmann (Eds.). *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und Naturwissenschaftliche Bildung am Ende der Schullaufbahn. Vol. 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 340-351). Opladen, Germany: Leske + Buderich.

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

Bishop, J., & Wößmann, L. (2004). Institutional effects in a simple model of educational production. *Education Economics, 12*(1), 17-38.

Block, R. Klein, E.D., Ackeren, I. van & Kühn, S.M. (2011). Leistungseffekte des Zentralabiturs? Eine kritische Auseinandersetzung mit bildungsökonomischen Interpretationen zu den Effekten der Prüfungsorganisation auf der Basis von PISA E 2003-Daten [Do statewide Abitur Examinations effect student achievement? A critical discussion of educational economics interpretations on the effects of different exit examination procedures on the basis of PISA-E 2003 data]. *bildungsforschung, 8*(1), 215-238.

Center on Education Policy (CEP) (2007). *"It's Different Now": How Exit Exams Are Affecting Teaching and Learning in Jackson and Austin.* Washington, D.C.

Center on Education Policy (CEP) (2011). *State High School Tests: Changes in State Policies and the Impact of the College and Career Readiness Movement.* Washington, D.C.

Cosentino de Cohen, C. (2010). *Examination Regimes and Student Achievement* (Dissertation). Princeton University, Princeton.

DeBray, E. (2005). A Comprehensive High School and a Shift in New York State Policy: A Study of Early Implementation. *The High School Journal*, 89(1), 18-45.

European Commission (2011): Germany. In European Commission. *Eurypedia. The European Encyclopedia on National Education Systems*. Retrieved January 29, 2012 from, https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Germany:Overview

Firestone, W.A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*(2), 95-113.

Fuchs, T., & Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics 32*(2), 433-464.

Goertz, M. E. & Massell, D. (2005). *Holding High Hopes: How High Schools Respond to State Accountability Policies* (CPRE Policy Briefs).

Hanushek, E. A., & Wößmann, L. (2007). *The role of education quality in economic growth*. Washington: World Bank.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Holme, J.J.; Richards, M.P.; Jimerson, J.B. & Cohen, R.W. (2010). Assessing the Effects of High School Exit Examinations. *Review of Educational Research*, 80(4), 476-526.

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

Klein, E. D. & Ackeren, I. van (in press). Challenges and Problems for Research in the Field of Statewide Exams. A Stock Taking of Differing Procedures and Standardization Levels. *Studies in Educational Evaluation*

Klein, D.; Kühn, S.; Ackeren, I. van & Block, R. (2009). Wie zentral sind zentrale Prüfungen? Zentrale Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich [State-wide Exit Exams in a national and international comparative analysis]. *Zeitschrift für Pädagogik*, 55(4), 596-621.

Kromrey, H. (1998): *Empirische Sozialforschung* [Empirical social research]. Stuttgart, Germany: UTB.

Kühn, S. (in press). Exploring the Use of Statewide Exit Exams to Spread Innovation – The Example of *Context* in Science Tasks from an International Comparative Perspective. *Studies in Educational Evaluation*

Kühn, S. (2010). *Steuerung und Innovation durch Abschlussprüfungen*? [Educational governance and innovation through exit examinations?] Wiesbaden, Germany: VS.

Lebart, L., Morineau, A., & Piron, M. (2000). *Statistique exploratoire multidimensionelle* [Multidimensional exploratory statistics]. Paris, France: Edition Dunod.

Longo, C. (2010). Fostering Creativity or Teaching to the Test? Implications of State Testing on the Delivery of Science Instruction. *Clearing House, 83*(2), 54-57.

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

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 the implementation of statewide exit examinations on teaching in advanced courses in high school]. *Unterrichtswissenschaft, 38*(2), 173-192.

Maag Merki, K., Klieme, E., & Holmeier, M. (2008). Unterrichtsgestaltung unter den Bedingungen zentraler Abiturprüfungen. Differenzielle Analysen auf Schulebene mittels Latent Class Analysen [Instructional practice under the conditions of central exit exams. Differential analyses at school level based on latent class analysis]. *Zeitschrift für Pädagogik,* 54(6), 791-808.

National Institute for Public Education. (2003). *The system of content regulation in Hungary: Public policy analysis*. Retrieved [1-25-2012] from, ftp://ftp.oki.hu/english/Content_Regulation.pdf

Nichols, S. L., & Berliner, D. C. (Eds.). (2005). *Collateral damage. How high-stakes testing corrupts America's schools*. Boston: Harvard University Press.

PISA-Konsortium Deutschland (2008). *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich. Zusammenfassung.* [PISA 2006 in Germany. Competencies of adolescents in the third Bundesland comparison] Retrieved January 29, 2012 from http://pisa.ipn.uni-kiel.de/Zusfsg_PISA2006_national.pdf

Sadler, D.R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, *30*(2), 175-194.

Schümer, G., & Weiß, M. (2008). *Bildungsökonomie und Qualität der Schulbildung. Kommentar zur bildungsökonomischen Auswertung von Daten aus internationalen Schulleistungsstudien* [The economics of education and the quality of schooling: Commentary on the educational economic analysis of data from international studies of student performance]. Frankfurt am Main, Germany: GEW.

Volante, L. 2007. Evaluating test-based accountability systems: An international perspective. *Proceedings of the Annual Meeting of the Association for Educational Assessment.* Stockholm, Sweden: AEA Europe.

Witten, E., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Mateo: Morgan Kaufman.

Wößmann, L. (2003). Central Exams as the "Currency" of School Systems: International Evidence on the Complementarity of School Autonomy and Central Exams. *DICE Report – Journal for Institutional Comparisons, 1*(4), 46-56.

Wößmann, L. (2008). Zentrale Abschlussprüfungen und Schülerleistungen [Central exit examinations and student achievement]. *Zeitschrift für Pädagogik, 54*(6), 810-826.

Appendix A

Table 4.
Indicators Discriminating Between Baden-Württemberg and Rhineland-Palatinate in the Abitur Subject of Mathematics (Advanced Level) According to the Students' Responses

Construct	**Indicator**	**V-test score (Reference BW)**	**Baden-Württenberg**		**Rhineland-Palatinate**					**Comparison: North Rhine-Westphalia**	
			M	*SD*	*M*	*SD*	η^2	ICC state in %	ICC school in %	*M*	*SD*
Overwhelming demands	Q1 (strongly disagree – strongly agree)	+6.6	3.14	.798	2.36	.871	.141	27.0	8.3	2.47	.865
Instructional approach	Q2 (very often – never)	+6.2	3.47	.757	2.78	.956	.120	(28.5)	10.6	3.08	.969
Overwhelming demands	Q3 (strongly disagree – strongly agree)	+4.7	3.41	.742	2.91	.865	.071	(10.5)	7.0	2.93	.795
Overwhelming demands	Q4 (strongly disagree – strongly agree)	+4.3	2.37	.888	1.87	.722	.059	(8.1)	6.9	1.89	.810
Instructional approach	Q5 (very often – never)	+4.2	2.19	.862	1.72	.740	.055	13.1	0.0	2.03	.818
Working on lesson content	Q6 (disagree – agree)	-4.2	1.78	.971	2.34	1.10	.055	(12.0)	6.8	2.05	1.01
Cooperative learning	Q7 (disagree – agree)	-4.3	1.57	.844	2.10	1.06	.060	(13.3)	10.9	1.97	1.05
Working on lesson content	Q8 (disagree – agree)	-4.5	1.89	.922	2.44	.946	.062	13.4	0.9	2.36	.935
Private tuition	Q9 (yes – no)	-4.7	1.57	.494	1.86	.338	.069	(13.1)*	3.3**	1.85	.354
Cooperative learning	Q10 (disagree – agree)	-5.0	1.66	.840	2.28	1.02	.081	18.4	3.7	1.89	.946
Working on lesson content	Q11 (disagree – agree)	-5.1	1.40	.616	1.91	1.02	.080	(16.6)	4.0	1.69	.752
Cooperative learning	Q12 (disagree – agree)	-5.5	1.47	.737	2.09	.939	.099	20.5	4.4	1.77	.897
Cooperative learning	Q13 (disagree – agree)	-5.8	1.65	.813	2.38	1.05	.111	24.9	3.6	1.91	.930
Homework	Q14 (after every lesson – never)	-6.6	1.18	.554	1.77	.831	.136	29.1	0.7	1.63	.964
Cooperative learning	Q15 (disagree – agree)	-6.9	1.45	.672	2.21	.997	.154	33.6	3.8	1.75	.872

* Rho=0.118 ** Rho=0.093

Table 5.
Non-subject Specific Indicators Discriminating Between Baden-Württemberg and Rhineland-Palatinate According to the Students' Responses

Construct	**Indicator**	**V-test score (Reference BW)**	**Baden-Württenberg**		**Rhineland-Palatinate**					**Comparison: North Rhine-Westphalia**	
			M	*SD*	*M*	*SD*	η^2	ICC state in %	ICC school in %	*M*	*SD*
Exam preparation outside the classroom	Q1 (strongly disagree – strongly agree)	+10.6	3.29	.813	2.29	.972	.238	38.3	0.0	3.42	.823
Exam preparation outside the classroom	Q2 (strongly disagree – strongly agree)	+8.3	2.19	1.15	1.35	1.78	.149	23.9	1.3	1.62	.994
Reasons for the choice of Abitur subject	Q3 (strongly disagree – strongly agree)	+8.0	2.16	.926	1.55	.681	.122	22.3	3.5	1.71	.695
Exam preparation outside the classroom	Q4 (strongly disagree – strongly agree)	+5.3	3.00	.936	2.51	1.01	.061	(8.8)	3.6	2.54	1.07
Expectancies with respect to exam preparation	Q5 (disagree – agree)	+5.2	2.97	.770	2.58	.796	.058	(8.9)	3.3	2.94	.856
Exam preparation	Q6 (yes – no)	-10.1	1.40	.532	2.01	1.16	.210	33.7*	4.1**	1.66	.642
Heterogeneity of students in the course	Q7 (yes – no)	-12.3	1.10	.307	1.66	.473	.330	48.4***	4.2****	1.11	.313

*Rho- 0.256 **Rho= 0.178 ***Rho= 0.387 ****Rho= 0.376

Table 6.
Non-subject Specific Indicators Discriminating Between Baden-Württemberg and Rhineland-Palatinate According to the Teachers' Responses

Construct	**Indicator**	**V-test score (Reference BW)**	**Baden-Württenberg**		**Rhineland-Palatinate**					**Comparison: North Rhine-Westphalia**	
			M	*SD*	*M*	*SD*	η^2	ICC state in %	ICC school in %	*M*	*SD*
Preferred organization of the statewide Abitur	Q1 (disagree – agree)	+6.8	2.88	1.17	1.43	.813	.343	51.6	0.6	1.24	.605
Preparation for the Abitur exam in lessons	Q2 (strongly disagree – strongly agree)	+4.6	3.90	.292	3.17	.996	.163	27.8	3.6	3.86	.347
Role of parents	Q3 (disagree – agree)	+4.0	2.41	.867	1.85	.666	.115	21.5	4.2	1.96	.778
Preparation for the Abitur exam in lessons	Q4 (strongly disagree – strongly agree)	+3.8	3.77	.418	3.19	.969	.108	19.3	0.0	3.86	.341
Pedagogical goals	Q5 (unimportant – very important)	+3.4	3.48	.659	3.01	.786	.087	15.5	0.0	2.98	.781
Role perception of teachers and students	Q6 (disagree – agree)	+3.3	2.92	.696	2.48	.718	.085	15.0	0.0	2.96	.565
Discussion of the results of the written Abitur exam	Q7(residual category) (agree – disagree)	-2.9	1.61	.486	1.84	.365	.065	(10.0)	2.0	1.63	.487
Heterogeneous course attendance (students who will/will not take an Abitur exam in the subject)	Q8 (yes, very much so – no, not at all)	-3.0	2.29	.654	2.75	.433	.141	(18.3)	8.0	2.19	.657

Table 6 (continued).

Construct	**Indicator**	**V-test score (Reference BW)**	**Baden-Württenberg**		**Rhineland-Palatinate**					**Comparison: North Rhine-Westphalia**	
			M	*SD*	*M*	*SD*	η^2	ICC state in %	ICC school in %	*M*	*SD*
Organization of teaching staff	Q9 (strongly disagree – strongly agree)	-3.2	1.92	.703	2.36	.800	.071	(12.5)	1.7	2.20	.792
Grading in advanced courses	Q10 (strongly disagree – strongly agree)	-3.9	2.40	.800	2.95	.701	.115	20.4	0.0	2.66	.805
Grading in basic courses	Q11 (strongly disagree – strongly agree)	-4.0	2.39	.730	2.91	.764	.097	(12.7)	11.0	2.81	.675
Discussion of the results of the written Abitur exam	Q12 (agree – disagree)	-4.1	1.74	.436	1.97	.158	.124	(22.2)	5.0	1.86	.347
Grading in basic courses	Q13 (strongly disagree – strongly agree)	-4.2	2.32	.820	2.87	.710	.108	(19.2)	2.4	2.68	.695
Preferred organization of the statewide Abitur	Q14 (disagree – agree)	-4.5	2.19	1.00	3.10	1.03	.159	36.2	2.1	3.08	.975
Grading in advanced-level courses	Q15 (strongly disagree – strongly agree)	-4.7	2.28	.776	2.98	.778	.164	(19.5)	10.6	2.85	.760
Heterogeneity of students in the course	Q16 (yes – no)	-4.8	1.25	.436	1.67	.468	.165	28.7*	2.4**	1.10	.312

*Rho= 0.198 **Rho= 0.177

Table 7.
Changes Brought About by the Introduction of the Statewide Abitur (in the Order of Amount of Agreement from Academic-Track Teachers in NW)

Construct	**Indicator**	**Direction of change / Nominations in %** (items answered on a 4-point Likert-scale)	
Narrowing of the delivered curriculum	Narrower focus of classroom instruction (subject 1)	*agree somewhat, agree*	
		2 years before the exam	76.8%
		1 year before	86.6%
		6 months	90.9%
		3 months	97.7%
	Use of classroom time for topics directly preparing students for the written Abitur exam (3 months before the exam)	*increase in the proportion of time*	
		6 months before the exam	65.1%
		3 months before	71.4%
Deprofessionalization	Competencies of the teacher reduced by the statewide Abitur (subject 1)	*a little / a lot*	73.6%
Experience of stress	I feel under greater pressure since the statewide Abitur has been introduced.	*agree somewhat, agree*	66.7%
School climate	Change in role perception of students and teachers	*now more like allies*	65.9%
Communication / Cooperation	Discussion of the results of written Abitur exams with colleagues	*increase in time allocation*	63.8%
	Teacher cooperation	*intensification*	60.8%
	Collaboration among teaching staff		52.1%
Internal management of the school	Use of exam results by teaching staff for purposes of school development and instructional development	*intensification of use*	55.3%
Student orientation in instruction	Basing the choice of topics covered in lessons on student interests	*less responsive to students' interests*	62%
Time gain / Reduction in workload	The statewide Abitur has relieved teachers of some workload.	*agree somewhat, agree*	56.6%
	Time has been gained in the context of preparing and conducting the Abitur exam	*yes, to some extent*	51.1%

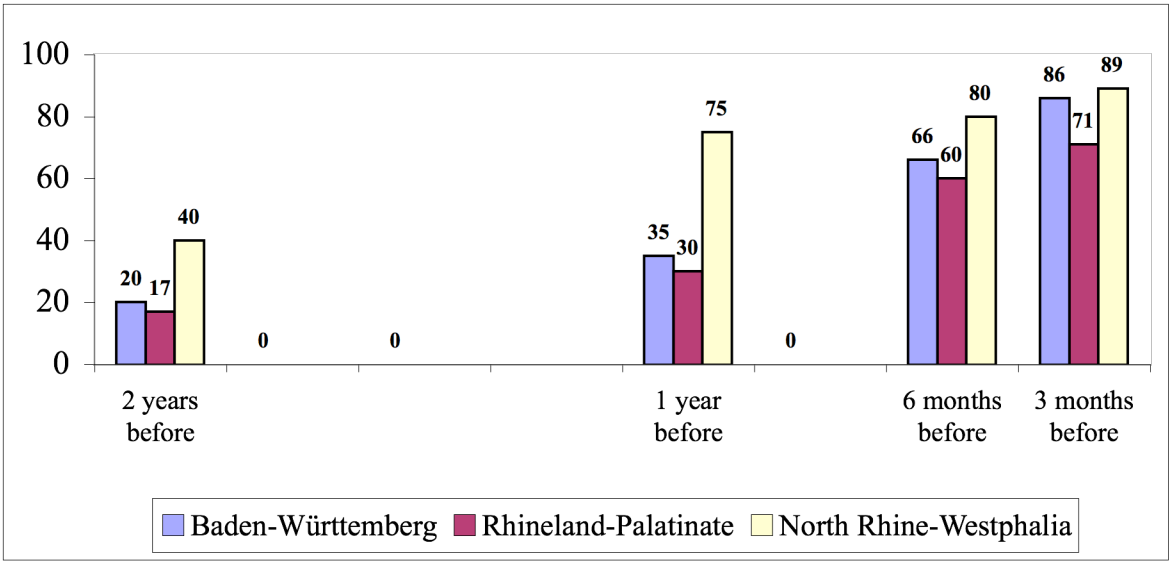

Figure 1. Proportion of Teachers Who Use 60% or More of Classroom Time for Topics Directly Preparing Students for the Abitur Exam (by Time and State)

Appendix B

Survey tools: student questionnaire, *Indicators Discriminating Between Baden-Württemberg and Rhineland-Palatinate in the Abitur Subject of Mathematics, Advanced Level* (see table 4)

	Item
Q 1)	The pace of instruction is so high that many students have difficulty keeping up.
Q 2)	Students work in groups.
Q 3)	If a student misses a few days, he or she has to struggle to catch up.
Q 4)	We have trouble keeping up with our homework.
Q 5)	The teacher and students discuss problems together.
Q 6)	In lessons, we often discuss approaches or solutions that we have developed in groups.
Q 7)	We often do group work (3–6 students).
Q 8)	In lessons, the teacher often does not say immediately whether an answer is right or wrong.
Q 9)	Paid for private tuition in an Abitur subject in the past 2 years.
Q10)	In lessons, we learn how to work effectively with a partner.
Q11)	In lessons, we often decide together with the teacher which topics to cover.
Q12)	In lessons, we learn what's important for effective classroom discussions.
Q13)	In lessons, we learn how to work together with others to everybody's benefit.
Q14)	Frequency of homework
Q15)	In lessons, we learn how to work effectively in groups.

Survey tools: student questionnaire, *Non-subject Specific Indicators Discriminating between Baden-Württemberg and Rhineland-Palatinate* (see table 5)

	Item
Q 1)	Looked at the questions set in past papers
Q 2)	Attended private preparation courses
Q 3)	Particular interest in the content of my Abitur subjects
Q 4)	Bought past papers from a bookshop
Q 5)	Teachers prefer to narrow the delivered curriculum
Q 6)	Aware of the topics covered in the Abitur exams
Q 7)	Courses attended by students who will/will not take an Abitur exam in the subject

Survey tools: teacher questionnaire, *Non-subject Specific Indicators Discriminating Between Baden-Württemberg and Rhineland-Palatinate* (see table 6)

Nr.	Item
Q 1)	Every written Abitur paper should be marked by a third examiner.
Q 2)	Extensive discussion in lessons of topics that might come up in the exam.
Q 3)	I have observed parental attempts to exert influence on instructional practice.
Q 4)	I have explained how students can best prepare for the Abitur exams.
Q 5)	Striving for absolute objectivity.
Q 6)	Students see their teachers more as allies in preparing for the challenges of the Abitur exam.
Q 7)	Discussion of the results of the written Abitur exam in another context than those named (residual category)
Q 8)	Heterogeneity of course attendance as a problem for instructional practice.
Q 9)	We use non-teaching hours to work together.
Q10)	When I'm giving grades, I consider whether a student has performed well or poorly relative to his or her own previous performance.
Q11)	When I'm giving grades, I consider how well a student has performed relative to the class as a whole.
Q12)	Discussion of the results of the written Abitur exam with other schools.
Q13)	When I'm giving grades, I consider whether a student has performed well or poorly relative to his/her own previous performance.
Q14)	Questions set in written exams should always be piloted before use.
Q15)	When I'm giving grades, I consider how well a student has performed relative to the class as a whole.
Q 16)	Courses attended by students who will/will not take an Abitur exam in the subject

About the Authors

Prof. Dr. Isabell van Ackeren
University of Duisburg-Essen, Faculty of Educational Sciences, Institute of Education

Email: isabell.van-ackeren@uni-due.de

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 research.

Dr. Rainer Block
Little Scientists' House, Berlin

E-Mail: rainer.block@haus-der-kleinen-forscher.de

Dr. Rainer Block is a scientific referee at the Little Scientists' House in Berlin (Germany). He is working in a project to encourage the enthusiasm of *three-to ten-year-old* boys and girls in natural science and technological phenomena. Before, he has coordinated the research project *Conditions and Effects of School-Based and Statewide Exit Exams* at University of Duisburg-Essen.

Esther Dominique Klein
Research Group & Graduate School *Teaching and Learning of Science*

Email: dominique.klein@uni-due.de

Esther Dominique Klein is a PhD-student at the research group and graduate school *Teaching and Learning of Science* (NWU) at the University of Duisburg-Essen, Germany. She is currently coordinating the research project *Forms and Effects of State-wide Exit Examinations in an International Comparison*. Dominique's main research interests are in educational governance, school system and international comparative education research.

Dr. Svenja Mareike Kühn
University of Duisburg-Essen, Faculty of Educational Sciences, Institute of Education

Email: svenja.kuehn@uni-due.de

Dr. Svenja Mareike Kühn is a postdoctoral research associate at the Faculty of Educational Sciences, University of Duisburg-Essen (Germany). Her thesis (2009) is titled "Analysis of Tasks in State-wide and School-based Exit Examinations. How are they Affected by Administrative Regulations?". Currently, her main research areas are in school system, education policy and educational governance research.

education policy analysis archives

Volume 20 Number 8 March 20, 2012 ISSN 1068-2341