Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 8 (2016), No. 2, 31–59

Esther Dominique Klein

How do teachers prepare their students for statewide exit exams?

A comparison of Finland, Ireland, and the Netherlands

Abstract

Statewide exit exams are often believed to have a positive impact on school effectiveness and the alignment between instructional practice and state standards because of their mandatory nature and the stakes attached for students and teachers. They may also, however, lead to teaching to the test and to a perceived de-professionalization of the teaching role. While some studies suggest a narrowing of contents and an increase in teacher-centered instruction, little is known about how the impact on instructional practices and teacher cognitions varies between different exam systems. This study compares the strategies teachers use to prepare their students for the exams at the end of upper secondary education in Finland, Ireland, and the Netherlands using a standardized questionnaire survey with responses from 385 teachers. The goal was to develop hypotheses about the relationship between differences in the exam procedures and the stakes attached, and the differences in teacher preparation strategies. The results suggest country-specific variations regarding teacher beliefs as to how much time should be spent on exam preparation; however, there were smaller differences in the strategies applied. Regression analyses indicated that the way in which preparation intensity was associated with the stakes for students and schools, and the attitudes towards the exams themselves varied across the three countries. The different exam systems appeared to affect preparation in markedly different ways, but nevertheless led to the exercise of comparable strategies.

Keywords

Statewide exit exams; International comparison; Teacher motivation; Teaching to the test

Dr. Esther Dominique Klein, Institute of Education, Faculty of Educational Sciences, University of Duisburg-Essen, Universitätsstr. 11, 45117 Essen, Germany
e-mail: dominique.klein@uni-due.de

Esther Dominique Klein

Wie bereiten Lehrkräfte ihre Schülerinnen und Schüler auf zentrale Abschlussprüfungen vor?

Ein Vergleich von Finnland, Irland und den Niederlanden

Zusammenfassung

Von zentralen Abschlussprüfungen wird aufgrund ihres verpflichtenden Charakters vielfach eine positive Wirkung auf die Effektivität von Schulen und die Kohärenz zwischen Unterrichtspraxis und staatlichen Standards erwartet. Demgegenüber stehen Befürchtungen, dass die Prüfungen z. B. vermehrtes teaching to the test bewirken könnten. Gleichwohl ist bislang kaum bekannt, welche Wirkungen unterschiedliche Prüfungsverfahren insbesondere auf Lehrkräfte entfalten können. Diese Studie vergleicht Prüfungsvorbereitungsstrategien von Lehrkräften der Sekundarstufe II in Finnland, Irland und den Niederlanden auf Basis einer standardisierten Befragung von 385 Lehrkräften. Ziel der Studie ist es, Hypothesen über den Zusammenhang zwischen Prüfungsverfahren und den damit verknüpften Konsequenzen einerseits und differenziellen Vorbereitungsstrategien von Lehrkräften andererseits herauszuarbeiten. Die Ergebnisse deuten darauf hin, dass der von den Lehrkräften für optimal gehaltene Zeitraum zur Vorbereitung länderspezifisch variiert, während sich nur geringe Unterschiede in den genutzten Vorbereitungsstrategien feststellen lassen. Regressionsanalysen deuten zudem darauf hin, dass der Zusammenhang zwischen Vorbereitungsintensität und Konsequenzen für Schüler/innen und Lehrkräfte sowie Überzeugungen der Lehrkräfte über die drei Länder hinweg unterschiedlich ausfällt. Die verschiedenen Prüfungssysteme beeinflussen die Lehrkräfte also auf unterschiedliche Weise, führen aber zu sehr vergleichbaren Strategien.

Schlagworte

Zentrale Abschlussprüfungen; Internationaler Vergleich; Lehrermotivation; Teaching to the test

1. Introduction

Statewide exit exams (SWEE) have specific and sometimes diverging or competing functions within school systems. SWEE are used to assure comparable, objective, and trustworthy certification and selection (e.g., Eckstein & Noah, 1993). In addition, they are often accredited with a positive effect on school effectiveness and student achievements (e.g., Bishop & Wößmann, 2004). The extent to which SWEE can actually affect and standardize the outcome of schooling, however, has not yet been clarified. Studies that investigate the association between graduation requirements and performance levels, performance variance, or achievement gains have varying results across different states, exam formats, subjects, course types, and stages of education (e.g., Baumert & Watermann, 2000; Bishop, 1995; Cosentino

de Cohen, 2010; Holme, Richards, Jimerson, & Cohen 2010; Shuster, 2012). Evidently, SWEE do not affect the outcomes of schooling in a direct and uniform way. However, they have an indirect influence on teaching and learning processes.

High stakes testing (HST) research in the USA suggests that HST has a limited positive impact on the quality of instruction, but can have tremendous side effects on organizational features (e.g., reallocation of educational resources), teacher cognitions (e.g., increased stress), and teaching and learning habits (e.g., teaching to the test). This is especially the case with schools that face challenging circumstances (e.g., Au, 2007). These findings, however, are not directly comparable with SWEE especially in Europe, since HST often differs from SWEE in content and organization (e.g., minimum competences vs. end-of-course exams) and the stakes attached (SWEE usually have high stakes attached for the students, but often only limited if any consequences for the schools) (Klein & van Ackeren, 2011). The effects of SWEE on instructional processes and the features of the exams that affect schools across different exam systems have so far only been investigated to a very limited degree and in specific regions. Moreover, SWEE studies usually explore one or two very similar SWEE systems so that it is difficult to identify the extent to which the observed effects are valid in other contexts. In addition, instructional processes are often studied shortly after a new exam system has been implemented, making it difficult to differentiate implementation effects from the long-term effects of SWEE (Klein, Krüger, Kühn, & van Ackeren, 2014).

This paper reports the results of an exploratory study that analyzed the strategies used by teachers to prepare their students for the upcoming exams in three European countries with SWEE systems that were introduced several decades ago. The SWEE systems differ in terms of design and the stakes attached for students and schools. In the study, teachers in upper secondary education in Finland, Ireland, and the Netherlands were asked about their preparation strategies in a standardized questionnaire survey. The goal was to shed light on the association between teachers' in-class preparation strategies and their attitudes towards the exams (i.e., motivation, confidence, and perceived de-professionalization) in the three different SWEE systems. The study outlines how the exams may affect instructional processes in theory and summarize the current state of research on SWEE. The study design and methodology are described as well as the SWEE systems in the three countries. Finally, the results of the study are reported.

2. Conceptual framework

SWEE are often believed to influence both the contents that are taught in schools and the way in which they are taught. SWEE therefore: prompt schools to try to ensure that students meet achievement standards through a positive *backwash* effect (Prodromou, 1995) that aligns the delivered and the intended curriculum (e.g., Eckstein & Noah, 1993); help the state emphasize core contents; support a

quick implementation of curricular innovations (e.g., Kühn, 2011); and may even influence didactic decisions and establish teaching standards in schools (e.g., van Ackeren, Block, Klein, & Kühn, 2012).

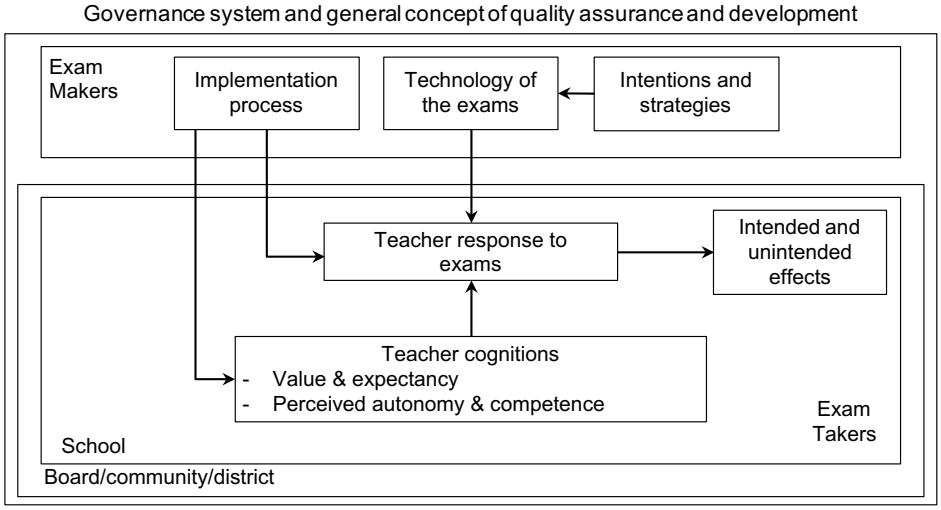
The theory of action assumed here can be explained with a model of SWEE effects developed by Bishop and Wößmann (2004), which assumes that teachers have little reason to maintain high standards in the classroom if there is no external incentive. In this context, SWEE results give the state and parents insight into the quality of the teachers' work, and a means to sanction poor work. Since teachers cannot lower requirements or change content to gain better results, they are forced to set high standards and focus their efforts on student learning (Bishop & Wößmann, 2004).

The model describes a relationship between SWEE and teacher responses, which is contingent on the specific design of the SWEE at the macro level (incentives, underlying governance intentions, and accompanying instruments) and the individual responses of teachers at the micro level, which are influenced by individual cognitions. The way in which changes in these two variables influence the relationship between SWEE and teacher responses is discussed in the following paragraphs.

2.1 SWEE systems

SWEE can be conceptualized as instruments of education policy (e.g., Eckstein & Noah, 1993). Modern political theories assume that such instruments do not affect schooling in a unilateral way, but are contingent on other idiosyncratic actors in the system. A theory that describes these instruments and processes while considering the influence of other actors in the education system is pending (Berkemeyer, 2010); therefore, this study uses the perspective of Educational Governance (EG) as an analytical lens. EG analyzes the collective capacity to act in a system that is shaped by multiple actors who face each other in specific constellations in a multilevel education system with functionally differentiated levels. The way in which teachers respond to an impulse from the state depends on the impulse itself and how they adapt it to their particular situation of the teacher. In turn, governance instruments are adapted to the anticipated response of school actors. Accordingly, the actors are connected through interdependencies which they try to manage with different instruments (Altrichter & Maag Merki, 2010). We can assume that the technology of the SWEE (procedures and tasks) and the process of implementation (support and control structures, and incentives) are part of and thus designed to fit a general strategy with which the state tries to manage its interdependency with schools. Both influence how teachers actually respond to the SWEE (i.e., how the SWEE are interpreted by school actors and what strategies they derive from them) and are thus decisive for whether the SWEE actually lead to the intended teacher behavior, or rather have unintended effects (see Figure 1).

Figure 1: Assumed relationship between macro and micro level factors (adapted from Klein, 2013, p. 147, based in Visscher, 2002, p. 66)



SWEE can be part of multilayered quality assurance in which they are one instrument among many to ensure that schools meet the common goal (e.g., the Netherlands). Alternatively, they can be highly atomized systems in which they help the state retain at least some standardization (e.g., Finland), or systems in which they are limited to their certification function, and are not meant to affect schools directly (e.g., Ireland). SWEE can contain diverging incentives, such as purposive incentives (e.g., teachers want students to pass the exams because that is the purpose of their job), material incentives (e.g., merit pay by exam results), or solidary incentives (e.g., good exam results can be used for profiling) (Clark & Wilson, 1961). The potency of these incentive constellations varies across countries. The relevance of SWEE results for university entrance may be extremely high or comparatively low. Authorities may or may not publish exam results of individual schools. Material incentives are rare in Europe but do play a role in some countries (e.g., the Netherlands). Likewise, SWEE procedures are adapted to their function, and thus differ regarding the subjects assessed, task format, task development, supervision of test-taking and marking procedures (Klein & van Ackeren, 2011). This variation is also likely to trigger differing responses from teachers:

- The incentives and standardization of the procedures may affect the perceived pressure.
- The capacity to deal with the SWEE may be contingent on additional support systems.
- Differences in the standardization will leave teachers varying room to act and let their own logic of action influence the process.

Research findings support the assumption that whether and to what extent teachers perceive pressure from SWEE depends on the incentives that accompany them. In Germany, where the (newly implemented) SWEE had incentives for students alone, teachers did not perceive them as overtaxing once they had gotten used to them (Maué, Maag Merki, & Oerke, 2012; Oerke, 2012). In studies from the USA and England, however, teachers perceived the high stakes for students as a motivating factor, but felt pressured by incentives for schools (Goertz & Massell, 2005; Massell, Goertz, Christensen, & Goldwasser, 2005; Perryman, Ball, Maguire, & Braun, 2011; Sipple, Killeen, & Monk, 2004; Zhang, 2009).

2.2 Teacher cognitions

The model by Bishop and Wößmann (2004) has a rather behaviorist perspective in which behavior modification is sought using rewards and sanctions. This excludes individual teacher cognitions. Therefore, in line with HST studies (e.g., Ryan & Weinstein, 2009), I argue that teacher cognitions can be explained using expectancy value models and self-determination theory.

With reference to expectancy value models, we can assume that the way in which teachers respond to SWEE is affected (a) by the attainment value, utility, and intrinsic value that the incentives have for a particular teacher, in relation to their cost; and (b) by the teacher's expectation that he or she can actually affect the outcome that is linked to the incentive (in this case, to make sure that all students pass the exams in the best possible way). Both value and expectancy vary between different incentives. The success or failure of students may have a different value for teachers than the prospect of the school being shut down. Moreover, they do not depend on an objective estimate of the task, but on the teacher's subjective cognitions including self-efficacy, causal attributions, and perceived difficulty (Eccles & Wigfield, 2002). For instance, if teachers believe that success in the exams is merely a result of cramming, the SWEE will not motivate them to raise standards.

External incentives may also affect and compromise existing intrinsic teacher motivation (e.g., concern for the welfare of students). The restrictive nature of SWEE might lead to a perceived lack of autonomy as the exams challenge the professionalism of teachers and their status as experts for instruction and assessment (Runté, 1998). With reference to self-determination theory (e.g., Gagné & Deci, 2005), it is likely that in a very restrictive (i.e., standardized) exam system that limits the professional autonomy of teachers and their opportunities to experience competence (e.g., in very low-performing schools) and which puts teachers under pressure to perform, teachers tend to use more controlling, teacher-centered instructional methods that lead to short-term instead of long-term improvement (Ryan & Weinstein, 2009). Unintended effects of SWEE may be that teachers limit content and methodology due to perceived pressure, not expecting to be able to affect exam outcomes in other ways; teachers may also feel stripped of their au-

tonomy. Teacher cognitions and resulting behavior may differ in the context of diverging impulses from the SWEE. Differences in the standardization of SWEE procedures affect the perceived autonomy and self-determination of the teacher; differences in the stakes attached to success lead to perceived pressure and can alter the value exam success has for teachers. The type and format of SWEE tasks also have a bearing on whether teachers expect improved results from better teaching or mere cramming.

3. Literature review

3.1 Effects on content

Among the unintended effects of SWEE often described in the research literature are that teachers narrow the delivered curriculum to declared or traditional exam content. In order to prepare their students for the contents of the exams, they may teach students in a way that lacks breadth or depth, or fails to consider the connections between different areas of content (*content approach*, Allalouf & Ben-Shakhar, 1998). Teachers may also feel less flexibility when it comes to incorporating current topics or issues of local interest; they may marginalize student interests. Especially when there are no standardized curricula, Schools may also reallocate time to those subjects that are part of the SWEE at the expense of time for independent study and other subjects.

A number of questionnaire surveys have investigated the differences in instructional practice in German states that switched from school-based to statewide exams. These studies indicated that at least in some cases teachers perceived the exams as restricting the room they had for individual support measures (Eickelmann, Kahnert, Lorenz, & Bos, 2011; Kühn & Racherbäumer, 2013; Racherbäumer & Kühn, 2013), limiting their ability to consider different perspectives, to discuss content critically, and to consider current topics or student wishes (Eickelmann et al., 2011). A longitudinal study in two German states showed that restrictions regarding the range of topics persisted over a longer period after the exams had been introduced (Jäger, Maag Merki, Oerke, & Holmeier, 2012; Maag Merki & Holmeier, 2008; Oerke, Maag Merki, Maué, & Jäger, 2013). Teachers who had low individual self-efficacy and felt uncertain regarding the requirements of the SWEE, or who thought that the collective self-efficacy in their school was low, were more prone to narrow the content taught (Jäger, 2012; Jäger et al., 2012). The way in which students perceived support from their teachers, however, did not differ between exam types (Holmeier & Maag Merki, 2012).

3.2 Effects on teaching and learning techniques

The exams may also have a backwash effect on the teaching and learning techniques used in the classroom, which may be adapted to familiarize students with the exam format (*familiarity approach*, Allalouf & Ben-Shakhar, 1998). This backwash effect is positive if the exams are *worth teaching to*; exam tasks that are less challenging or do not require hands-on activities or self-initiated and creative solutions may have the trade-off of an increased use of reproductive learning and a decrease in student-centered, creative and less controlling forms of learning.

A comparative questionnaire study from the USA (Vogler, 2006, 2008; Vogler & Carnes, 2009) found that teachers teaching courses that ended with a SWEE felt that their instruction was influenced by the stakes for the students; however, this did not lead to a more teacher-centered instruction than in courses that did not conclude in a SWEE. The ratio of teacher- and learner-oriented practice differed across subjects, but did not seem to depend on the associated stakes. An interview study by Krüger, Won, and Treagust (2013) in Australia revealed the use of more teacher-centered teaching techniques in the final two years before the exams. Other findings also suggested a higher use of controlling teaching techniques (as rated by the students) in the context of SWEE in Germany (van Ackeren et al., 2012); this varied across subjects, however.

Two other German studies indicated that SWEE also affected the deep structure of instruction. Elaborative learning techniques appeared to be higher in courses ending with a SWEE (Baumert & Watermann, 2000) and increased in courses where the SWEE was newly introduced (Maag Merki, 2011; Maag Merki & Holmeier, 2008; Maag Merki, Holmeier, Jäger, & Oerke, 2010). This relationship, too, varied across subjects and course types. In a study by Maag Merki, increased elaboration techniques could be found for mathematics and English, but not for biology or German (Maag Merki, 2011). Baumert and Watermann (2000), on the other hand, found increased elaboration strategies in mathematics and advanced physics courses. However, this study did not control for context and individual factors. Moreover, elaboration strategies also seemed to depend on how the exams were handled in different schools (Maag Merki, Klieme, & Holmeier, 2008).

4. Study design

While the reported findings suggest that classroom activities and preparation strategies are affected by SWEE, no studies have yet systematically investigated the influence of differences in the SWEE on processes and outcomes at the micro level. The study at hand therefore tries to shed light on the effects of SWEE on direct exam preparation strategies in three different SWEE systems at the end of upper secondary education (USE) in Finland (FI), Ireland (IR) and the Netherlands (NL). The study is part of a larger research project investigating the functions and effects

of SWEE (for a detailed description of the project, see Klein, 2013). The goal of this paper is to answer the following research questions:

1) How much time do teachers believe should be spent on teaching exam-relevant content?
2) What strategies do teachers use to prepare their students for the SWEE?
3) How do preparation strategies differ across subject areas?
4) How is the intensity of preparation affected by attitudes towards and cognitions in the context of the SWEE?

The study uses an international comparative approach in which the three countries are treated as individual cases that are described using the same methodology. The cases are then juxtaposed and compared. Because of the lack of empirical findings, the study is exploratory and follows the design of an external validation study in which hypotheses are not forwarded, but deduced from the results (van de Vijver & Leung, 1997).

4.1 The SWEE systems

The countries were chosen systematically in a least similar cases design based on their dissimilarity in two aspects.

School governance system: The aim was to choose three countries with heterogeneous governance structures based on a model by Schmid, Hafner and Pirolt (2007). NL was chosen because of its "pressure-and-support"-approach toward school governance, whereas FI represents a country in which decisions in the school system are based on consensus. In IR, the state is a traditionally weak player in the school system, whereas other actors have considerable influence on decision making. As described in Section 2.2, these differences go in hand with diverging intentions regarding the SWEE.

Standardization of the exams: Especially the German studies described above analyze the effects of SWEE with a very low standardization level (Klein & van Ackeren, 2011). The aim therefore was to include SWEE systems that differ in their standardization level, but have a higher standardization level than the German SWEE.

The characteristics of the countries are described in Table 1.

Table 1: Characteristics of the SWEE systems (as of 2011)

	FI	IR	NL
Type of Governance	Legitimacy type with local empowerment	School empowerment between Legitimacy and Bureaucracy type	Legitimacy and Efficiency type with school empowerment
		Exam procedures	
Name	*Ylioppilastutkinto*	*Leaving Certificate Examinations*	*Eindexamen vwo*
Range	SWEE only	SWEE only	SWEE & SBEE
No. of exams	4 (Choice from Finnish, Swedish, modern foreign language, mathematics, general studies)	around 7; only Irish is compulsory	All subjects in final exam; some only in SBEE
Supervision	School	External supervisor	School
Marking	Preliminary marking by teacher; main marking by central exam commission	External marker from central exam commission	First marking by teacher, second marking by teacher from other school
Marking standards	none	Standardized marking schemes	Standardized marking schemes
Overall standardization	moderate	very high	SWEE high / SBEE low
		Incentives	
Stakes for students	Passing SWEE necessary for university entrance Additional requirements possible	University places are allocated in a central system based on SWEE results only	SWEE and SBEE both count for 50% of final grade University entrance partly based on final grade
Stakes for schools	Media publishes results	None	Inspectorate publishes results and uses them in external evaluation

4.1.1 Exam procedures

Whereas the Finnish and Irish exams are completely external, the Dutch SWEE is complemented with a school-based exit exam (SBEE) that accounts for 50 % of the final grade in each subject. Dutch schools are free to organize the SBEE as they wish. In FI, students choose four exams from a pool of five subjects (Finnish, Swedish, another modern foreign language, mathematics, general studies). In IR, students are free to choose the subjects they take during USE and in the exams; only Irish is compulsory. In NL, students choose a study profile in addition to a core set of subjects in USE. They take a SWEE and a SBEE in most core and profile subjects.

The exam papers are developed by a central exam commission in all three countries. In FI and NL, the exams are supervised by school personnel. In IR external supervisors are used; in addition, the marking of papers is also carried out by an external examiner using a standardized marking scheme. In FI and NL, in contrast, teachers are part of the marking process. In FI, the teacher conducts a preliminary marking; the final and decisive marking is done by an examiner from the central exam commission. No official marking scheme is issued by the exam commission. In NL, the teacher marks the papers first on the basis of a standardized marking scheme; a teacher from another school makes a second marking. Accordingly, the chances for teachers to support their students (within legal means) in the supervision and marking (e.g., because they have a better understanding of the students' answers) are higher in FI and NL than in IR.

4.1.2 Exam tasks

In NL, the SWEE are intended to provide a reliable benchmark of student competencies across schools, whereas the SBEE assess performance using a more holistic approach. The SWEE tasks in all subjects therefore contain largely standardized test formats (e.g., short answers and multiple choice). In IR, too, the aim is to have a reliable exam standard. This is related to the high relevance the SWEE have for students (see below). Since there are no additional SBEE, the SWEE must be both reliable and valid regarding all content specified in the syllabi. Exams in languages and arts mostly contain questions where students have to produce short texts of their own, whereas in mathematics or science exams, students must solve problems in short answers. In FI, the curriculum for USE contains very broad specifications of content. Therefore, the exam papers usually follow the same layout each year with respect to the number of tasks and the sequence of contents. Depending on the subjects, the tasks usually ask candidates to produce short texts (e.g. in the languages) or solve problems. In addition, they may include questions regarding compulsory project work or interdisciplinary issues.

4.1.3 Stakes

In both FI and IR, the SWEE results are the sole measure used to decide whether or not students may go on to university. In IR, a central office allocates university places on the grounds of the SWEE results alone. In FI, the SWEE are a *sine qua non*, but universities can ask for additional requirements (e.g., results of internal assessments); thus, graduates can compensate for poor SWEE results. In NL, students can balance the SWEE results with their results in the SBEE: University entrance depends on the overall final grade. The formal pressure to ensure students obtain very good results is highest in IR and lowest in NL.

From the school perspective, the SWEE only have tangible stakes in NL, where the average results of schools are published and used by the inspectorate as a benchmark in external evaluation. Low-performing schools that do not improve over several years may face closure. In IR, the results of schools are not published. Although the inspectorate may consider results during inspection, there are no tangible consequences for schools. In FI, the state abstains completely from controlling its schools through exam data; however, municipalities can use the SWEE results for accountability and the media regularly publish school results. The pressure to have good SWEE results to avoid sanctions from the state or other stakeholders, or to maintain a good reputation, are formally higher in NL than in the other two countries; however the perceived solidary incentives may nevertheless be felt to a high degree in these countries. Table 1 summarizes the features of the three systems.

4.2 Data sources and methodology

The study targeted schools preparing for the SWEE at the end of USE in larger cities with more than 100,000 inhabitants. With the exception of schools with less than 200 students in IR and NL, all possible schools were asked to participate. In FI, 15 *Lukio* schools (about 25 % of the target population) participated. Thirteen post-primary schools participated in IR (about 10 % of the target population). In NL, the focus was broadened to smaller cities due to recruiting problems. Despite this effort, only seven schools participated; thus, less than 5 % of the target population was represented.

Moreover, the school sample was not representative because only two schools in both FI and IR, and no school in NL stated to be located in a deprived area. In addition, coeducational schools were underrepresented in the Irish sample.

4.2.1 Teacher sample

A standardized questionnaire was sent to 1,328 teachers in the selected schools in spring, 2011. The survey was intended to comprise all teachers in USE; as the questionnaire was voluntary, the response rates were low (see Table 2).

Table 2: Average response rates (fully completed questionnaires)

		Overall	Finland	Ireland	Netherlands
Online	% by no. of announced teachers	19.4 %	24.2 %	6.1 %[a]	28.0 %
	% by total no. of teachers at school[b]		35.1 %		
Paper/Pencil	% by no. of announced teachers	33.7 %	43.9 %	29.6 %	27.5 %[c]

[a] Only one school. [b] Two schools where the number of teachers participating was not announced before. [c] Schools partly announced participation rates that were much lower than total number of possible participants.

38 returned questionnaires (mostly online) were not completed fully. Of the 385 fully completed questionnaires, 30 were excluded because teachers did not teach a subject that concluded in a SWEE (N = 355; N_{FI} = 133; N_{IR} = 167; N_{NL} = 56). An additional questionnaire from IR was excluded because it was not completed in earnest. The majority of teachers were female in FI (69.7 %) and IR (73.3 %), but not in NL (42.9 %). This roughly reflects the actual ratio within secondary schools in these countries. In all three countries, more than 80 % of the teachers had work experience of at least six years.

The teachers were asked to name their first or main subject and focus on that subject throughout the questionnaire. In all three countries, the majority of teachers (between 44.4 % and 55.4 %) taught a subject from the languages and arts area (L/A). Mathematics or science subjects (M/Sc) were taught by between 24.7 % and 32.3 % of the respondent teachers. The remaining teachers taught humanities and other subjects ending with a SWEE.[1]

4.2.2 Methodology

To limit cross-cultural item bias, the questionnaire was translated into Finnish, English, and Dutch using forward and backward translation and bilingual experts (van de Vijver & Leung, 1997). However, differential response styles across cultures may result in extreme ratings in one culture and more conservative ratings in another (method bias). One way to test for method bias is to collect additional data for validation. Since this was not possible here, the data were analyzed separately for each country; patterns and correlations within the data of the countries were compared instead of concrete scores.

To analyze teacher attitudes toward exam preparation, teachers were asked how much of the instructional time during USE they thought should be predominantly used for exam-relevant content. Options were “throughout USE”, “in the last school year”, “in the last six months”, “in the last four to six weeks”, and “never”. In IR,

1 In the Dutch sample, no teachers remained in the “others” category after non-SWEE subjects were excluded.

the option "Transition Year" (an optional year before the regular USE) was included.

To investigate the strategies teachers use to prepare their students for the exams, they were asked to indicate which of the following approaches they used for direct exam preparation during USE (5-point Likert scale, 1 = never to 5 = very often):

1) I focus on content that is often dealt with in the exams.
2) I prepare students for the format of the exams specifically.
3) I discuss the typical assignments and how they should be conceived with the students.
4) I coach my students in answer formats that are often used in the exams (e.g., multiple choice).
5) I discuss singularities of the exams with my students that they have not come to know in my classes.

The first item addresses the content approach, the other four items the familiarity approach (Allalouf & Ben-Shakhar, 1998). All items loaded on the same factor and were merged in a scale indicating the intensity of exam preparation (Cronbach's $\alpha = .77/\alpha_{FI} = .80$; $\alpha_{IR} = .73$; $\alpha_{NL} = .79$).

To investigate how the preparation intensity is associated with attitudes and cognitions in the context of the SWEE, multiple regressions with ordinary least squares and different predictors at the individual level were calculated separately for each country. Because the sample sizes were too small for multi-level analyses, the degree to which teachers were nested within schools could not be accounted for.

The attitudes and cognitions of teachers were modelled with the following variables, based on theoretical considerations regarding the motivation theories described above (all scales were developed by the author and measured on a 5-point Likert scale):

1) Attitudes towards the exams:
   - *Perceived utility/benefit of the exam papers* ($\alpha = .90/\alpha_{FI} = .89$; $\alpha_{IR} = .91$; $\alpha_{NL} = .90$; five items, e.g., "The fact that the learning objectives of the syllabus are broken down in the exam tasks helps me to decide which topics and issues should be especially focused on in my lessons");
   - *Motivation by solidary incentives*[2] ($\alpha = .83/\alpha_{FI} = .85$; $\alpha_{IR} = .82$; $\alpha_{NL} = .76$; three items, e.g., "I am motivated to prepare my students optimally for the exams because our school can distinguish itself with good exam results"):
   - *Motivation by purposive incentives* (Individual item: "I am motivated to prepare my students optimally for the exams because my students' prospects very much depend on their exam results");

2 Because there are no material incentives in FI and IR, this type was not considered.

2) Teacher cognitions:
   - *Confidence in exam preparation* (α = .65/α_{FI} = .73; α_{IR} = .68; α_{NL} = .34[3]; three items; e.g., "I often have the feeling that my classes prepare students poorly for the exams"; reversed item);
   - *Perceived deprofessionalization by SWEE* (α = .76/α_{FI} = .74; α_{IR} = .64; α_{NL} = .85; four items; e.g., "I hardly have any chance to carry out my own ideas regarding subject content because I have to conform to the content of the exams");
   - Demographics: *subject area* (reference L/A)*, sex, seniority* (15 years or more vs. less than 15 years).

The analyses were run with IBM© SPSS in a forced stepwise approach. First, the intensity was regressed on the demographics only. Secondly, the attitudes towards the SWEE were included. The final model contained all predictors.

Model assumptions were checked with scatter plots, histograms, and the Kolmogorov-Smirnov test (Urban & Mayerl, 2006). Multicollinearity was checked using the variance inflation factor and casewise diagnostics were used to detect outliers (Field, 2009). The results of additional regression analyses without outliers largely resembled the results with the full sample; therefore, they were not excluded. Missing data were excluded listwise.

5. Results

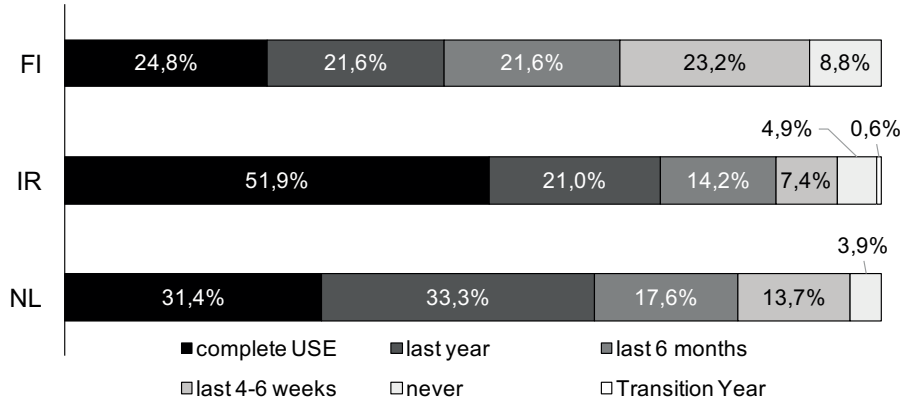
5.1 Time teachers believe should be used for teaching exam-relevant content

The time teachers believed should be spent on exam-relevant content was investigated first (see Figure 2). The category "complete USE" indicated the strongest focus on exam-relevant contents. In NL, 31.4 % agreed with this category; 24.8 % of the Finnish teachers agreed with this category and, in IR, this option was chosen by more than half the teachers (51.9 %). In contrast, 23.2 % of the teachers in FI, but only 13.7 % of the Dutch and 7.4 % of the Irish teachers thought that limiting the focus on exam-relevant contents to the last four to six weeks of USE was an agreeable strategy. The results indicate that in IR, teachers were more likely to focus on exam-relevant content for a longer period of time than in NL and FI especially. The results, however, do not permit conclusions regarding the substantive orientation or intensity of the exam preparation within this period.

There were no significant differences between teachers from different subject areas in any of the countries.

3 The confidence scale had no satisfactory internal consistency for NL. An auxiliary confidence index was used instead. This had only two items that had both the same high factor loading of .85, indicating a high validity, and a satisfactory internal consistency (α = .60).

Figure 2: Teachers' perception of the amount of time that should be predominantly spent on exam contents in upper secondary education (in %)



N_{FI} = 125; N_{IR} = 162; N_{NL} = 51

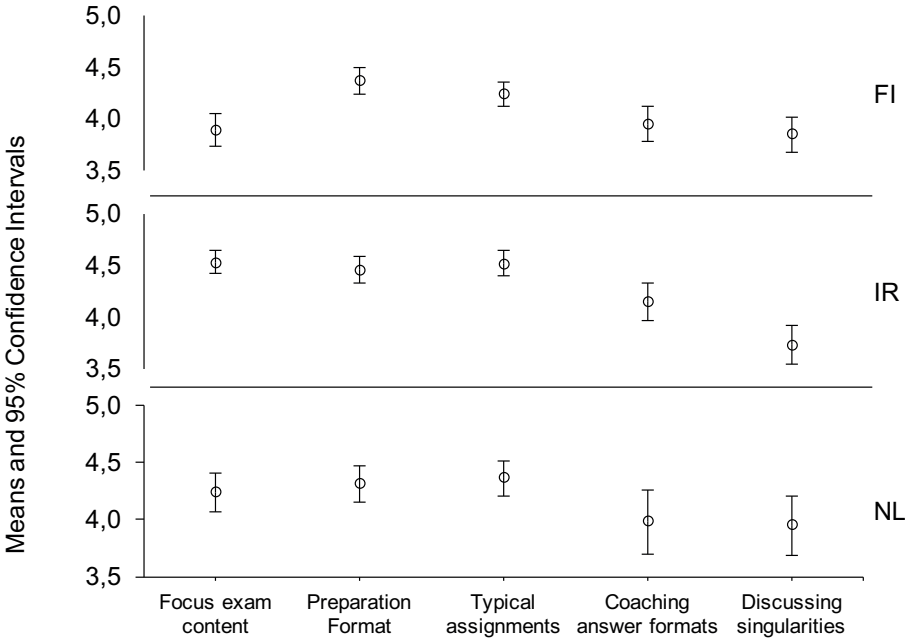
5.2 Preparation strategies

Use of all preparation strategies was well above the theoretical mean. The factor structure of the preparation intensity scale, which integrates these five items, does not suggest a difference between the content and the familiarity approach. In principal axis factors analyses, all items loaded on the same factor explained between 52 % and 59 % of the variance. Accordingly, it is unlikely that teachers differentiate between content and familiarity approaches.

The scores of the three countries are illustrated by the dots in Figure 3. In all three countries, the rating was highest for familiarizing the students with the type and form of the SWEE and discussing typical exam assignments (second and third items). Coaching answer formats and discussing singularities, on the other hand, received comparatively low ratings. Moreover, the 95 % confidence intervals (CI) were broader for these two items.

In IR and NL, teachers seemed to focus their preparation on typical exam content more often, indicating that teachers in these countries were more focused on the contents of the SWEE at least during the phase of direct exam preparation than in FI; here, the content item received the second lowest rating of all five strategies.

Figure 3: Preparation strategies of teachers by country



Note. 5-point Likert scale (1 = never to 5 = very often; values 2 through 4 not labelled)

The last two strategies are particularly likely to be affected by individual teaching habits as well as by subject-specific testing cultures. Therefore, the preparation strategies were also compared by subject groups.

5.3 Subject-specific differences in preparation strategies

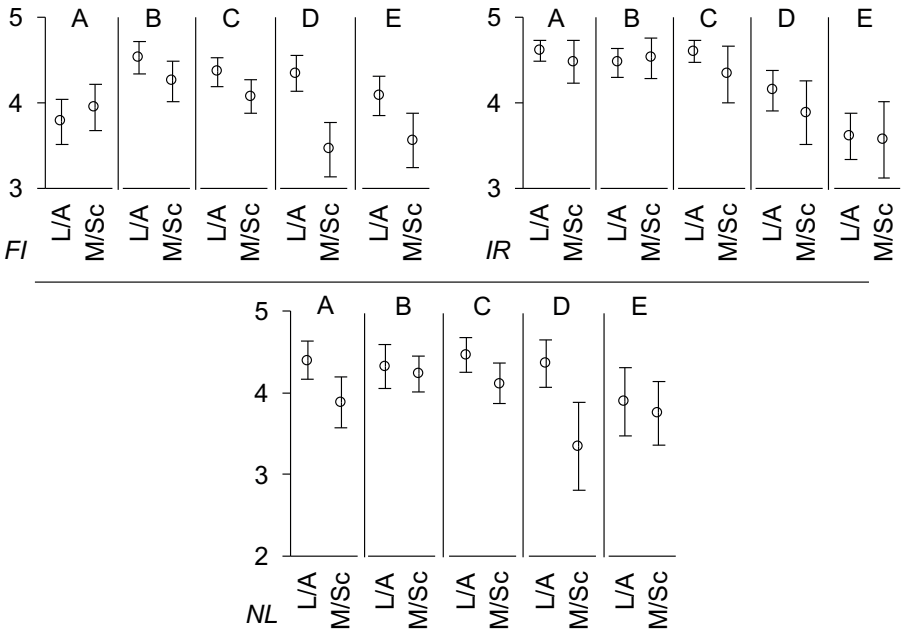
Because there were few humanities teachers and teachers in the "others" category, only the L/A and M/Sc teachers were juxtaposed. The comparison showed that in all three countries, L/A teachers reported a nominally higher degree of agreement to most strategies than M/Sc teachers; however, the comparison revealed differences in the subject patterns within countries. In FI, for instance, there were no differences between M/Sc and L/A teachers regarding the focus on typical content. The preparation for the format was higher for L/A teachers (M = 4.33, SD = 0.71) than for M/Sc teachers (M = 4.24, SD = 0.44), t (95) = 1.78, d = 0.36; likewise, the focus on typical exam assignments was higher for L/A teachers (M = 4.40, SD = 0.62), than for M/Sc teachers (M = 3.88, SD = 0.60), t (95) = 2.09, d = 0.43. This might be because the curricula for science subjects are very open in FI; however, the 95 % CI overlap considerably (see Figure 4). On the other hand, L/A teachers reported much more frequent coaching of specific answer formats (M = 4.37,

SD = 0.76) than M/Sc teachers (*M* = 3.45, *SD* = 1.06), *t* (94) = 5.0, *d* = 1.01, and a more frequent discussion of the singularities of the exams (*M* = 3.90, *SD* = 1.12) than M/Sc teachers (*M* = 3.76, *SD* = 0.75), *t* (92) = 2.98, *d* = 0.61). Here, the CI showed no overlap, but were far broader than for the other items.

The Finnish result was mirrored by the Dutch result; here, too, the L/A teachers reported more frequent coaching of answering formats (*M* = 4.15, *SD* = 1.09) than M/Sc teachers (*M* = 3.89, *SD* = 1.13), *t* (45) = 3.80, *d* = 1.10. L/A teachers focused on typical content more often (*M* = 4.61, *SD* = 0.54) than M/Sc teachers (*M* = 4.48, *SD* = 0.78), *t* (34)[4] = 2.81, *d* = 0.85. Here, the CI did not overlap.

In IR, the effect sizes did not suggest any differences between L/A and M/Sc teachers.

Figure 4: Preparation strategies of language or arts teachers (L/A) and mathematics or science teachers (MSc) by country (means and 95% confidence intervals)



Note. A = focus typical contents; B = preparation for format; C = typical assignments; D = coaching answer formats; E = discussing singularities; 5-point Likert scale (1 = never to 5 = very often; values 2 through 4 not labelled).

4 Levene's test indicated unequal variances (*F* = 4.67, *p* = .036); as a result, degrees of freedom were adjusted from 45 to 34.

5.4 Influence of attitudes and cognitions on preparation intensity

Next, aspects that affect the intensity of preparation (henceforth, intensity) were analyzed with regression analyses. Table 3 illustrates the means and CI of the outcome variable and the continuous predictor variables by country.

Table 3: Statistics for the predictor variables by country

		Intensity of preparation	*Utility of exam papers*	*Solidary incentives*	*Purposive incentives*	*De-professionalization*	*Confidence*
FI	*M (SE)*	4.08 (.06)	3.59 (.072)	3.27 (.09)	3.94 (.08)	2.29 (.07)	3.04 (.07)
	CI_{95}	[3.97, 4.19]	[3.45, 3.73]	[3.09, 3.45]	[3.79, 4.09]	[2.14, 2.43]	[2.90, 3.18]
IR	*M (SE)*	4.30 (.05)	4.02 (.06)	3.56 (.08)	4.49 (.06)	3.01 (.07)	3.04 (.08)
	CI_{95}	[4.20, 4.39]	[3.89, 4.14]	[3.39, 3.72]	[4.38, 4.61]	[2.88, 3.15]	[2.91, 3.17]
NL	*M (SE)*	4.17 (.08)	3.97 (.09)	3.64 (.10)	3.95 (.12)	2.33 (.12)	*4.36*[a] *(.08)*
	CI_{95}	[4.01, 4.32]	[3.80, 4.15]	[3.43, 3.85]	[3.70, 4.19]	[2.09, 2.57]	*[4.21, 4.51]*

Note. CI_{95} = 95 % confidence interval; 5-point Likert scales.
[a] Auxiliary confidence scale in NL.

With regard to FI, the regression analysis confirmed the results of the preceding sections, which showed that M/Sc teachers reported a lower intensity than L/A teachers; this was also true when seniority and sex, cognitions, and attitudes were taken into account (see Table 4). Accordingly, the subject area effect was not mediated by subject-specific attitudes towards the exams. Because the CI for this regression coefficient did not contain 0, the precision of the point estimate is given.

Model 2 showed that the attitudes toward SWEE only partly affected the intensity. While a higher motivation by both types of SWEE incentives entailed an increased intensity, there was no significant association between intensity and the perceived utility of the exam papers. Regarding the incentives, the beta value of the motivation caused by the stakes for students was a little higher than that of the motivation caused by profiling opportunities; here, the CI of the regression coefficient contains 0, indicating that the actual precision of the point estimate is low.

Finally, Model 3 showed that teacher cognitions were not associated with intensity. Teachers who felt more restricted in terms of their professional autonomy did not report a higher intensity. Interestingly, the confidence teachers had in their exam strategies was not related to the intensity. Altogether, the final model only explained about 12 % (adj. R^2) of the variance, which indicates that the intensity was affected by other factors not considered in the model.

Table 4: Regression of intensity on predictors (FI)

	M1[a]					M2					M3				
				CI_{95} (*B*)					CI_{95} (*B*)					CI_{95} (*B*)	
Predictors	*B*	*SE*	ß	*LL*	*UL*	*B*	*SE*	ß	*LL*	*UL*	*B*	*SE*	ß	*LL*	*UL*
(constant)	4.32	.10				3.13	.37				3.04	.53			
15y or more	-.12	.11	-.09	-.34	.11	-.19	.11	-.15^{+}	-.41	.03	-.21	.12	-.17+	-.44	.02
Male	-.08	.13	-.05	-.33	.18	-.06	.12	-.04	-.30	.18	-.05	.13	-.04	-.30	.20
M/Sc	-.33	.13	-.26*	-.59	-.07	-.32	.13	-.24*	-.57	-.07	-.34	.13	-.26*	-.60	-.08
Hum	-.16	.16	-.09	-.48	.17	-.07	.16	-.04	-.39	.24	-.05	.16	-.03	-.37	.27
Oth	-.56	.26	-.19*	-1.08	-.032	-.43	.26	-.15^{+}	-.93	.08	-.41	.26	-.14	-.94	.11
Papers						.08	.07	.10	-.06	.21	.07	.07	.09	-.07	.21
Solidary						.10	.05	.17^{+}	-.00	.21	.11	.06	.18^{+}	-.00	.22
Purposive						.17	.07	.22*	.03	.30	.17	.07	.22*	.03	.31
Confidence											.05	.07	.06	-.17	.11
Deprof											-.03	.07	-.04	-.09	.19
Goodness of fit	R^2 = .09 *adj.* R^2 = .05 $F(5, 121)$ = 2.31*					R^2 = .19 *adj.* R^2 = .13 $F(8, 116)$ = 3.32**					R^2 = .19 *adj.* R^2 = .12 $F(10, 113)$ = 2.72**				

Note. CI_{95} = 95 % confidence interval, LL = lower limit, UL = upper limit; ** $p < .01$; * $p < .05$; $^{+}$ $p < .10$
[a] For M1, residuals were not normally distributed. A log-transformation of the outcome variable did not improve this, so that results must be interpreted with caution.

With regard to IR, the preceding section had not revealed any significant differences between the subject areas (see Table 5). The regression analyses confirmed this even when other demographics or teacher attitudes towards the exams were held constant. (In the category "others", which was very small, the CI contains ß = 0 and thus suggests that the precision of the estimate was low.) In addition, the regression did not display an association between the intensity and the seniority of the teachers or their sex.

Regarding teacher perception of the exams, the results were contradictory to those in FI. Whereas the motivation gained from purposive and solidary incentives was not associated with the intensity, a more positive view of the utility of exam papers entailed a higher intensity, even though the association was not very strong (ß = .17 in Model 2). Model 3 additionally showed that confidence and perceived deprofessionalization were not related to the intensity.

The model that best fit the data was the second; however this explained only 6 % (adj. R^2) of the variance. Accordingly, the variance in the intensity was probably due to other factors not considered in the model. Moreover, this result might even be overestimated because in all three models, the residuals were not normally distributed. This could not be improved by transforming the outcome variable.

Table 5: Regression of intensity on predictors (IR)

	M1					M2					M3				
				CI_{95} (B)					CI_{95} (B)					CI_{95} (B)	
Predictors	*B*	*SE*	ß	*LL*	*UL*	*B*	*SE*	ß	*LL*	*UL*	*B*	*SE*	ß	*LL*	*UL*
(constant)	4.45	.16				3.40	.41				3.25	.53			
15y or more	-.00	.10	-.00	-.20	.19	-.02	.10	-.01	-.21	.18	-.02	.10	-.02	-.23	.18
Male	-.11	.11	-.08	-.34	.11	-.09	.11	-.06	-.30	.13	-.08	.11	-.06	-.30	.14
M/Sc	-.10	.12	-.07	-.34	.15	-.12	.12	-.09	-.37	.12	-.11	.12	-.08	-.35	.14
Hum	.16	.13	.11	-.09	.41	.15	.13	.10	-.10	.40	.18	.13	.12	-.08	.43
Oth	-.42	.23	-.15^{+}	-.87	.03	-.36	.23	-.13	-.81	.08	-.33	.23	-.12	-.79	.12
Papers						.13	.06	.17*	.02	.25	.13	.06	.17 *	.01	.25
Solidary						.06	.05	.10	-.04	.16	.06	.05	.10	-.04	.16
Purposive						.07	.07	.08	-.08	.21	.07	.07	.08	-.07	.21
Confidence											.01	.06	.01	-.11	.13
Deprof											.02	.06	.03	-.10	.15
	R^2 = .05 *adj.* R^2 = .02 $F(5, 156)$ = 1.64 (ns)					R^2 = .10 *adj.* R^2 = .06 $F(8, 151)$ = 2.18*					R^2 = .11 *adj.* R^2 = .04 $F(10, 146)$ = 1.71^{+}				

Note. CI_{95} = 95 % confidence interval, LL = lower limit, UL = upper limit; * $p < .05$; $^{+}$ $p < .10$.

With regard to NL, the first model confirmed the subject differences observed earlier. Moreover, the analyses showed that male teachers appeared to report a lower intensity than female teachers. Model 2 indicated that the difference by sex increased, but differences by subject area decreased, when the attitudes towards the exams were held constant.

An increased motivation by solidary and purposive incentives did not alter the intensity significantly (Model 2). Including the confidence and deprofessionalization in Model 3 increased the standardized regression coefficient of the motivation by solidary incentives (ß = .23) and purposive incentives (ß = -.11). However, for both predictors, the CI contained ß = 0, indicating that the actual precision of the point estimate was low.

Model 3 explained 24 %, and thus a higher proportion of the variance in the intensity than in the other two countries; however, because the Dutch sample was rather small, the determination coefficient might be overestimated (Urban & Mayerl, 2006).

Table 6: Regression of intensity on predictors (NL)

	M1					M2					M3				
				CI_{95} (B)					CI_{95} (B)					CI_{95} (B)	
Predictors	*B*	*SE*	ß	*LL*	*UL*	*B*	*SE*	ß	*LL*	*UL*	*B*	*SE*	ß	*LL*	*UL*
(constant)	4.77	.25				3.32	.67				3.22	.68			
15y or more	.03	.16	.03	-.28	.35	-.06	.15	-.05	-.36	.25	-.05	.16	-.04	-.36	.27
Male	-.34	.16	-.29^{+}	-.66	-.02	-.39	.16	-.33*	-.72	-.06	-.40	.17	-.34*	-.73	-.06
M/Sc	-.34	.17	-.27^{+}	-.68	.01	-.24	.17	-.19	-.57	-10	-.25	.17	-.20	-.58	.09
Hum	.14	.23	.08	-.32	.59	.11	.21	.06	-.33	.54	.10	.22	.06	-.33	.54
Papers						.30	.11	.34*	.08	.53	.30	.12	.34*	.07	.53
Solidary						.15	.11	.20	-.07	.36	.17	.11	.23	-.06	.40
Purposive						-.05	.10	-.07	-.25	.15	-.08	.11	-.11	-.29	.15
Confidence[a]											.10	.15	.10	-.21	.13
Deprof											-.04	.09	-.06	-.40	.19
	R^2 = .19 *adj.* R^2 = .13 $F(4, 50)$ = 2.99*					R^2 = .35 *adj.* R^2 = .25 $F(7, 47)$ = 3.58**					R^2 = .36 *adj.* R^2 = .24 $F(9, 45)$ = 2.86**				

Note. CI_{95} = 95 % confidence interval, LL = lower limit, UL = upper limit; ** $p < .01$; * $p < .05$; $^{+}$ $p < .10$.
[a] Auxiliary confidence scale.

6. Discussion

The preceding analyses have revealed commonalities and differences in the preparation strategies of teachers in the three countries.

First, teachers in Ireland were most likely to find it necessary to spend a lot of time in USE in exam preparation. The strategies applied, however, did not reveal a more intense focus on the exam preparation in IR than in FI or NL. Teachers in all three countries seemed to apply the strategies to a high degree (though the scores cannot be compared directly, see above). There was also no clear distinction between the content and the familiarity approach within the countries, although a qualitative difference existed between the country patterns with regard to the content approach item. This was rated lower than most other items in FI, but among the highest in IR and NL. This might signify differences caused by the curriculum. In FI with its very open curriculum and very standardized sequence of contents in the exam papers, exam-relevant content might in fact represent a narrow understanding of the contents that will appear in the exams, whereas in IR, the SWEE cover the whole syllabi, so that all contents are exam-relevant. The latter might also explain why Irish teachers believe that the majority of time in USE should be spent on exam-relevant content.

The factors affecting the intensity of preparation also diverged between the three countries. The regression did not show an association between the intensity

of preparation and motivation by the stakes for students or schools in IR and NL, but did so in FI. Here, motivation by exam stakes for the students was a stronger and positive predictor. While other studies have suggested that stakes for schools have a huge impact on teachers (Perryman et al., 2011; Zhang, 2009), the results of this study point in a different direction. While teachers felt highly motivated by purposive incentives, these did not have a strong influence, and especially not in IR, where the stakes for students were highest (this might also be a ceiling effect (see Table 2). Including the motivation by material incentives in NL might lead to a different result.

In addition, confidence in teachers' own exam preparation strategies did not appear to be related with the actual intensity in any of the three countries. This was also true for perceived deprofessionalization. This may mean that, despite theoretical considerations, preparation strategies are not connected to the attitudes and cognitions of teachers in the context of the exams. The data suggest that both confidence and deprofessionalization seem to be affected by the SWEE system (Klein, 2013). However, cognitions and motivation do not seem to affect teacher preparation strategies. The lack of connection between teacher cognitions and the strategies they apply reflects the findings made by Vogler (2006, 2008) and Vogler and Carnes (2009) at least to some degree. Both approaches, however, rely on questionnaire surveys, which are vulnerable to measurement errors; for instance, teachers who wish to express discontent with the SWEE system may exaggerate the perceived deprofessionalization or the intensity of preparation. Therefore, the results of this exploratory approach should be validated in further studies with instruments that are less contingent on subjective estimation by teachers (e.g., standardized classroom observation).

The utility of the exam papers, on the other hand, seemed to strongly influence the intensity in IR and NL, but not in FI. The fact that the actual exam papers seemed to play such a strong role for the intensity at least in these two countries confirms the backwash effect of the exams and underlines how important it is that exam tasks have a high quality. In FI, however, the papers appeared to be less important for the intensity. Other analyses in the context of the project revealed that exam papers were perceived to be helpful for very different purposes in the three countries. Regarding the breadth and depth of contents in the classroom, for instance, Finnish teachers perceived the exam papers as less helpful than their Irish and Dutch counterparts (see Klein, 2013, pp. 322-324). This may indicate that exam papers will have a stronger backwash effect on instruction when teachers perceive them as helpful for decisions they make in the classroom on the breadth and depth of content. It also indicates that the perceived helpfulness of the instrument has a stronger influence on preparation strategies than the incentives and the pressure teachers may feel.

Regarding subject-specific effects, the results revealed that the time item did not differ across subject areas, but that preparation strategies did. In addition, the juxtaposition showed that the subject patterns differed across countries, which was probably caused by characteristics of the SWEE. In IR, the patterns of higher and

lower agreement to the different strategies were largely the same for L/A and M/Sc teachers. Effects of subject culture and a distinctive uptake might have been covered because only subject areas were taken into account. This assumption is supported by the findings by Maag Merki (2011), which showed differences between subjects, but no clear distinction between subject areas. The lack of subject differences might also indicate that the high relevance of the exams leads to a focus on exam preparation by all teachers that overrides possible expectable differences due to subject culture, sequencing of content, or the range of possible content and task formats.

In NL, L/A teachers used the contents approach more often than M/Sc teachers. This would be an expected result as the range of possible content is probably much higher in L/A subjects than in M/Sc subjects. In FI and IR, however, the same difference could not be found. While this may merely reflect differences in the translation of the item, the results may also mirror the fact that the odds for science subjects to be chosen as exam subjects are smaller in FI and IR than in NL. In direct exam preparation, science teachers may therefore have a stronger focus on exam contents to compensate for a lesser focus on exam contents during USE due to the low ratio of exam candidates in their courses. This would support the findings in the Baumert and Watermann (2000) study, which suggested that differences between physics and mathematics (albeit in student performance) were actually differences between compulsory and voluntary subjects.

The strongest subject differences could be found in the coaching of answer formats, which indicates that this is very much contingent on the task format. In NL, for instance, the L/A subjects are assessed with short answers and multiple choice – formats that are probably used to a lesser extent in L/A classrooms.

Taken together, the findings of the study lead to the hypothesis that:

1) A more positive view of the utility of the exam papers, dependent on the format of exam papers and curricula, will lead to a stronger backwash effect.
2) The stakes for students and schools affect teacher motivation but do not influence their practice.
3) Perceived competence and deprofessionalization are affected by the SWEE system, but do not, in turn, affect the intensity of exam preparation.
4) The effects of SWEE on instructional processes are subject-specific.
5) Besides subject cultures, subject effects on exam preparation are context-specific (e.g., depend on the curriculum, exam papers, and compulsory subjects in different SWEE).

Nevertheless, only a small to medium proportion of the variance in the preparation intensity could be explained with the chosen predictors. Besides measurement errors, one explanation is that the variance was caused by other aspects at the micro level (competencies, general job satisfaction, etc.) or the meso level (teacher collaboration, school culture, social background and performance level of students, etc.). While this study can only generate hypotheses about relationships regarding the items included at micro level, additional research is needed to confirm these find-

ings and develop hypotheses about the meaning of other individual and collective factors.

7. Limitations

While the results of the study may lead to a better understanding of the SWEE systems and preparation strategies applied in schools, the results must be interpreted with some caution. First, the schools represented in the sample were mostly schools in advantageous contexts. In schools with challenging circumstances, preparation strategies might, for instance, be much higher to compensate for the less favorable conditions. It is also plausible that only very active and development-oriented schools or schools with a negative view of the exams were ready to participate in the study. The same bias might also apply for the teachers as they were not obliged to participate. Moreover, the majority of teachers had several years of experience with the exams; novice teachers might therefore report a higher preparation intensity that might be more influenced by their cognitions or attitudes.

Secondly, because of the study's exploratory nature, the items were phrased in a very open way (e.g., "singularities of exam tasks" were not further specified). This might have led to individual interpretations and thus distorted the results. The items were not phrased specifically for each subject; therefore "exam-specific content" may carry a different meaning in L/A than in M/Sc. The items also carry different meanings across the countries due to the differences in the breadth and depth of the underlying curricula (see Section 4.1.2). It is difficult to conclude whether the differences observed are actually caused by the SWEE or by a general heterogeneity between the different cultural settings that were observed. This, however, is a general problem in the analysis of effects at the macro level. Since it is hardly possible to analyze SWEE in a quasi-experimental design that controls for other factors, research can only try to disentangle the association between instructional practice and different aspects of the exam system through the comparison of different exam systems using a research design that follows Governance research approaches (Maag Merki, Langer, & Altrichter, 2014). Such a design should consider the interaction between actors at different levels of the school system through qualitative and quantitative multi-level analyses and completive qualitative case studies.

Thirdly, regarding the precision in parameter estimation in the regression analyses, the analyses suggested that the subject area, attitude towards exam papers, and the motivation (in part) seemed to have a statistically significant and relatively precise influence on preparation intensity. However, while the CI of these regression coefficients did not contain 0 in most cases, its lower bound usually came very close to 0. There is therefore some uncertainty as to how precise the assumed influence of these predictors really is.

To sum up, the data are limited regarding the generalizability of results into the complete population of schools in the respective countries. In addition, the data obviously do not allow for statements about preparation strategies in other countries with their own distinctive governance structures, exam systems, and cultural diversifications. For the goal of this exploratory study, which was to provide a first attempt to map the terrain, the approach taken was sufficient. In further studies that actually test hypotheses, the use of more distinctive and possibly subject-specific instruments should be discussed.

Acknowledgements

This paper was funded by the German Research Foundation (DFG). It was financially supported by the *Zentrum für empirische Bildungsforschung* at the University of Duisburg-Essen.

References

Ackeren, I. van, Block, R., Klein, E. D., & Kühn, S. M. (2012). The impact of statewide exit exams – A descriptive case study of three German states with differing low stakes exam regimes. *Education Policy Analysis Archives*, *20*(8), 1–28.

Allalouf, A. & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests, *Journal of Educational Measurement, 35*(1), 31–47.

Altrichter, H., & Maag Merki, K. (2010). Steuerung der Entwicklung des Schulwesens. In H. Altrichter & K. Maag Merki (Eds.), *Handbuch Neue Steuerung im Schulsystem* (pp. 15–39). Wiesbaden, Germany: VS.

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

Baumert, J., & Watermann, R. (2000). Standardisierung durch die Abiturprüfung – Zentralabitur oder dezentrale Prüfungsorganisation? In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie-mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. (2nd ed., pp. 341–351). Opladen, Germany: Leske + Budrich.

Berkemeyer, N. (2010). *Die Steuerung des Schulsystems. Theoretische und praktische Implikationen*. Wiesbaden, Germany: VS.

Bishop, J. H. (1995). The impact of curriculum-based external examinations on school priorities and student learning. *International Journal of Educational Research*, *23*(8), 653–752.

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

Clark, P. B., & Wilson, J. Q. (1961). Incentive systems – A theory of organizations. *Administrative Science Quarterly*, *6*(2), 129–166.

Cosentino de Cohen, C. (2010). *Examination regimes and student achievement* (Doctoral Dissertation). Retrieved from http://search.proquest.com/docview/305214076

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132.

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

Eickelmann, B., Kahnert, J., Lorenz, R., & Bos, W. (2011). Das Zentralabitur in Nordrhein-Westfalen aus der Lehrerperspektive – Veränderungen für den Unterricht. *SchulVerwaltung NRW*, *12/2011*, 31–32.

Field, A. P. (2009). *Discovering statistics using SPSS (and sex and drugs and rock'n' roll)* (3rd ed.). London, United Kingdom: Sage.

Gagné, M., & Deci, E. L. (2005). Self-Determination theory and work motivation. *Journal of Organizational Behavior*, *26*(4), 331–362.

Goertz, M. E. & Massell, D. (2005). *Holding high hopes – How high schools respond to state accountability policies* (CPRE Policy Briefs, RB 42). Philadelphia, PA: Consortium for Policy Research in Education. DOI: 10.12698/cpre.2005.rb42

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.

Holmeier, M., & Maag Merki, K. (2012). Unterstützung im Unterricht im Kontext der Einführung zentraler Abiturprüfungen. In K. Maag Merki (Ed.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (pp. 155–178). Wiesbaden, Germany: VS.

Jäger, D. J. (2012). Herausforderung Zentralabitur – Unterrichtsinhalte variieren und an Prüfungsthemen anpassen. In K. Maag Merki (Ed.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (pp. 179–205). Wiesbaden, Germany: VS.

Jäger, D. J., Maag Merki, K., Oerke, B., & Holmeier, M. (2012). Statewide low-stakes tests and a teaching to the test effect? An analysis of teacher survey data from two German states. *Assessment in Education: Principles, Policy & Practice, 19*(4), 451–467.

Klein, E. D. (2013). *Statewide exit exams, governance, and school development. An international comparison*. Münster, Germany: Waxmann.

Klein, E. D., & Ackeren, I. van (2011). Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels. *Studies in Educational Evaluation*, 37(4), 180–188.

Klein, E. D., Krüger, M., Kühn, S. M., & Ackeren, I. van (2014). Wirkungen zentraler Abschlussprüfungen im Mehrebenensystem Schule. Eine Zwischenbilanz internationaler und nationaler Befunde und Forschungsdesiderata. *Zeitschrift für Erziehungswissenschaft*, *17*(7), 7–33.

Krüger, M., Won, M., & Treagust, D. F. (2013). Teachers' perceptions on the changes in the curriculum and exit examinations for biology and human biology. *Australian Journal of Teacher Education*, *38*(3), 41–58.

Kühn, S. M. (2011). Exploring the use of statewide exit exams to spread innovation – The example of context-orientation in science tasks from an international comparative perspective. *Studies In Educational Evaluation*, *37*(4), 189–195.

Kühn, S. M., & Racherbäumer, K. (2013). Standardisierung und/oder Individualisierung? Empirische Befunde zur Umsetzung von Maßnahmen zur individuellen Förderung im Kontext zentraler Abschlussprüfungen. *Unterrichtswissenschaft*, *41*(2), 172–189.

Maag Merki, K. (2011). The introduction of state-wide exit examinations – Empirical effects on math and English teaching in German academically oriented secondary schools. In M. A. Pereyra, H.-G. Kotthoff, & R. Cowen (Eds.), *Pisa under examination. Changing knowledge, changing tests, and changing schools* (pp. 125–141). Rotterdam, Netherlands: Sense.

Maag Merki, K., & Holmeier, M. (2008). Die Implementation zentraler Abiturprüfungen – Erste Ergebnisse zu den Effekten der Einführung auf das schulische Handeln der Lehrpersonen. In E.-M. Lankes (Ed.), *Pädagogische Professionalität als Gegenstand empirischer Forschung* (pp. 233–244). Münster, Germany: 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. *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 Analysis. *Zeitschrift für Pädagogik, 54*(6), 791–808.

Maag Merki, K., Langer, R., & Altrichter, H. (Eds.). (2014). *Educational Governance als Forschungsperspektive – Strategien, Methoden, Ansätze*. Wiesbaden, Germany: VS.

Massell, D., Goertz, M. E., Christensen, G., & Goldwasser, M. (2005). The press from above, the pull from below – High school responses to external accountability. In B. Gross & M. E. Goertz (Eds.), *Holding high hopes. How high schools respond to state accountability policies* (pp. 17–41). CPRE Research Report Series. Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania.

Maué, E., Maag Merki, K., & Oerke, B. (2012). Emotionales Erleben des Zentralabiturs von Lehrpersonen in Bremen – Längerfristige Effekte der Implementation zentraler Prüfungen. In S. Hornberg & M. Parreira do Amaral (Eds.), *Deregulierung im Bildungswesen* (pp. 109–130). Münster, Germany: Waxmann.

Oerke, B. (2012). Emotionaler Umgang von Lehrkräften und Schüler/-innen mit dem Zentralabitur – Unsicherheit, Leistungsdruck und Leistungsattributionen. In K. Maag Merki (Ed.), *Zentralabitur. Die längsschnittliche Analyse der Prozesse und Wirkungen der Einführung zentraler Abiturprüfungen in zwei Bundesländern* (pp. 119–153). Wiesbaden, Germany: VS.

Oerke, B., Maag Merki, K., Maué, E., & Jäger, D. (2013). Zentralabitur und Themenvarianz im Unterricht – Lohnt sich teaching-to-the-test? In D. Bosse, F. Eberle, & B. Schneider-Taylor (Eds.), *Standardisierung in der gymnasialen Oberstufe* (pp. 27–49). Wiesbaden, Germany: VS.

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

Prodromou, L. (1995). The backwash effect – From testing to teaching. *ELT Journal*, *49*(1), 13–25.

Racherbäumer, K., & Kühn, S. M. (2013). Zentrale Prüfungen und individuelle Förderung. *Zeitschrift für Bildungsforschung*, *3*(1), 27–45.

Runté, R. (1998). The impact of centralized examinations on teacher professionalism. *Canadian Journal of Education*, *23*(2), 166–181.

Ryan, R. M., & Weinstein, N. (2009). Undermining quality teaching and learning – A self-determination theory perspective on high-stakes testing. *Theory and Research in Education*, *7*(2), 224–233.

Schmid, K., Hafner, H. & Pirolt, R. (2007). *Reform von Schulgovernance-Systemen. Vergleichende Analyse der Reformprozesse in Österreich und bei einigen PISA-Teilnehmerländern* (IBW-Schriftenreihe, Vol. 135). Wien, Austria: IBW.

Shuster, K. (2012). Re-Examining exit exams – New findings from the educational longitudinal study of 2002. *Education Policy Analysis Archives*, *20*(3), 1–31.

Sipple, J. W., Killeen, K., & Monk, D. H. (2004). Adoption and adaptation – School district responses to state imposed learning and graduation requirements. *Educational Evaluation and Policy Analysis*, *26*(2), 143–168.

Urban, D., & Mayerl, J. (2006). *Regressionsanalyse – Theorie, Technik und Anwendung* (2nd ed.). Wiesbaden, Germany: VS.

Vijver, F. van de, & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

Visscher, A. J. (2002). A Framework for Studying School Performance Feedback Systems. In A. J. Visscher & R. Coe (Eds.), *School Improvement Through Performance Feedback* (pp. 41–72). Lisse, NL: Swets & Zeitlinger.

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. (2008). Comparing the impact of accountability examinations on Mississippi and Tennessee social studies seachers' instructional practices. *Educational Assessment*, *13*(1), 1–32.

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 AERA, San Diego, CA.

Zhang, Y. (2009). *Conflicts between state policy and school practice – Learning from Arizona's experience with high school exam policies*. Washington, DC: Center on Education Policy.