International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: https://www.tandfonline.com/loi/tsed20

The ambiguous influence of high-stakes testing on science teaching in Sweden

Anders Jonsson & Lotta Leden

To cite this article: Anders Jonsson & Lotta Leden (2019): The ambiguous influence of high-stakes testing on science teaching in Sweden, International Journal of Science Education, DOI: 10.1080/09500693.2019.1647474

To link to this article: https://doi.org/10.1080/09500693.2019.1647474

Published online: 29 Jul 2019.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION
https://doi.org/10.1080/09500693.2019.1647474

Routledge
Taylor & Francis Group

The ambiguous influence of high-stakes testing on science teaching in Sweden

Anders Jonsson and Lotta Leden

Faculty of Education, Kristianstad University, Kristianstad, Sweden

ABSTRACT

Tests convey messages about what to teach and how to assess. Both of these dimensions may either broaden or become more uniform and narrow as a consequence of high-stakes testing. This study aimed to investigate how Swedish science teachers were influenced by national, high-stakes testing in science, specifically focusing on instances where teachers' pedagogical practices were broadened and/or narrowed. The research design is qualitative thematic analysis of focus group data, from group discussions with Swedish science teachers. The total sample consists of six teachers, who participated in 12 focus group discussion during three consecutive years. Findings suggest that the national tests influence teachers' pedagogical practice by being used as a substitute for the national curriculum. Since the teachers do not want their students to fail the tests, they implement new content that is introduced by the tests and thereby broaden their existing practice. However, when this new content is not seen as a legitimate part of teachers' established teaching traditions, the interpretation and implementation of this content may replicate the operationalisations made by the test developers, even though these operationalisations are restricted by demands for standardisation and reliable scoring. Consequently, the tests simultaneously broaden and narrow teachers' pedagogical practices.

ARTICLE HISTORY
Received 7 November 2018
Accepted 20 July 2019

KEYWORDS
Argumentation; laboratory work; summative assessment

Introduction

As expressed by Black and Wiliam (1998) more than 20 years ago, raising the standards of learning in school is an important national priority in most countries. Governments throughout the world have therefore been increasingly enthusiastic in making changes in pursuit of this aim, for instance by formulating national standards and mandating enhanced programs for external testing of student performance. Sweden is no exception and during the last decade a number of political initiatives, with a stronger focus on summative assessment, have been implemented, including a more detailed national curriculum (with more 'requirements' than goals for long-term learning), grading for younger students, and an increased weight given to test scores in relation to teachers' assessments.

CONTACT Anders Jonsson anders.jonsson@hkr.se

These changes have been made in spite of research suggesting that they may actually be counterproductive, for instance since low-achieving students tend to lower (rather than improve) their performance when being graded (e.g. Klapp, 2015). Furthermore, in their review of research on the impact of high-stakes testing on student motivation, Harlen and Deakin Crick (2003) conclude that results from such tests have been found to have a 'particularly strong and devastating impact' (p. 196) on low-achieving students. Nevertheless, as noted by for instance Black and Wiliam (1998) and Kohn (2000), politicians generally act as if external testing (or grading) will, on its own, improve student learning.

High-stakes testing has also been shown to have an impact on teachers and their teaching, which is the focus of this article. The relationship between high-stakes testing and classroom practice is, however, a complex matter. While the primary consequences of high-stakes testing is that curricular content is narrowed and subject area knowledge fragmented into test-related pieces (see e.g. Minarechová, 2012, for an overview), there are also studies showing that certain types of high-stakes tests may actually lead to curricular content expansion or have other positive consequences. As suggested in the meta-synthesis by Au (2007), a critical factor in this regard is the design of the tests. This is of particular interest in Sweden, since the national tests have a strong tradition of providing teachers with exemplary assessment tasks. This could, in turn, be expected to have positive consequences for science teaching, for instance by curricular content expansion and a broadening of teachers' repertoire of different assessment formats. Currently, however, it is not known how, or to what extent, Swedish science teachers' pedagogical practices[1] are influenced by these tests.

Background

High-stakes, external, and standardised testing

The focus of this article is on high-stakes, external, and standardised testing. A test is considered *high-stakes* when its results are used to make important decisions that affect students, teachers, administrators, communities, schools, and/or districts (Madaus, 1988; in Gunnemyr, 2011). A typical example is when students' grades are either determined or influenced by results from individual tests.

External tests are, according to Gunnemyr (2011, p. 48), tests where: (a) external actors have taken control, completely or in part, of one or more of the four sub-processes of the 'testing process' (i.e. test design, performance, assessment/scoring, and use of the results/outcome), and (b) where the individual teacher cannot refuse or object to how these sub-processes are implemented.

Finally, that tests are *standardised* means that all test takers have to do the same test and under the same circumstances, sometimes also at the same time. Furthermore, the test should be scored in the same way for all test takers. Standardisation is basically a way to secure comparability among the test takers. It is important to note that standardisation does not per se imply the use of only written format, selected-response items, or any other restrictions to the test design. Oral and practical performance may also be standardised (such as 'objective structured clinical examinations', or OSCE, in medical education). However, restrictions are often imposed for practical or economic reasons, such as time restrains, the cost of having assessors make observations of individual students, or that all students (despite their individual differences) need to be able to perform the same test.

Consequences of high-stakes testing

That high-stakes testing often has consequences for those affected is logical. If students' grades are influenced by their results on a particular test, they are likely to be motived to perform well on that test. Similarly, if teachers are affected by their students' results, they too have an incentive to 'teach to the test'. This steering effect is problematic primarily since any test can sample only a fraction of the full curriculum. To focus exclusively on what is covered on tests therefore involves a serious reduction of the subject content to be taught. It should be noted that this may apply equally well to the assessment format, since, for reasons explained above, selected-response and short answer items are normally over-represented on tests, while other assessment formats may be more suitable and/or prevalent as part of classroom assessment.

One of the most elaborate models for the influence of external tests on teachers' pedagogical practice has been proposed by Gunnemyr (2011), which takes into consideration both implicit and explicit communication from the test developers to the test users.[2] According to this model, there are three products that may convey messages, namely: (a) instructions on how to distribute and use the test and the test results, (b) the actual test, and (c) assessment/scoring instruments. While the message in (a) is explicit, both (b) and (c) contain implicit messages about, for instance, what kind of knowledge or skills that are important enough to be included in the test, how different levels of performance should be valued, and which items/task are appropriate for measuring/assessing this knowledge or skills. In line with what was noted above, this means that tests may convey messages about both *what to teach* (i.e. curricular content) and *how to assess* student performance in relation to this particular content. Below, we will therefore present research investigating these two dimensions separately, starting with consequences for curricular content.

Consequences of high-stakes testing on curricular content

The term 'curricular content', as used here, refers to any kind of knowledge or skills specified in a curriculum. The current Swedish national curriculum (Swedish National Agency for Education, 2018), for example, specifies certain 'abilities' that the students should be given the conditions to develop during their time in school. For the sciences, these 'abilities' are to:

- use knowledge in science to review information, communicate, and take a stand on issues that require knowledge in science,
- carry out systematic investigations, and
- use scientific concepts, models, and theories to describe and explain phenomena and relationships in the world around us.

The current Swedish national curriculum also stipulates core content to be taught for different subjects and grades. For example, during grades 4–6, teachers are required to teach the following core content to the students in biology:

- How mental and physical health are affected by sleep, diet, movement, social relationships and addictive substances. Some common diseases and how they can be prevented and treated (in relation to the category 'Body and health').

- People's dependence on and the impact on nature and what this means for sustainable development. Ecosystem services, such as decomposition, pollination, and purification of water and air (in relation to the category 'Nature and society'). (Swedish National Agency for Education, 2018, p. 168)

In total, there are 14 core content standards for biology during year 4–6, as well as a number of performance standards, linked to the 'abilities' above. As an example of the performance standards, students should be able to 'carry out simple field studies and other studies based on given plans and also *formulate* simple questions and plans which *after some reworking* can be systematically developed' (Swedish National Agency for Education, 2018, p. 172, emphasis in original).

In theory, high-stakes tests could, by sampling specific content, influence teachers to either include or exclude any kind of content from their teaching (i.e. related to either [broader] 'abilities' and/or [more limited] content knowledge). This means that tests could broaden as well as limit the content covered by teachers. For example, if the mandating authority would like to implement teaching of new content, they could make it part of the tests, which would then create a strong incentive for the teachers to make this content part of their teaching (since otherwise their students might fail the test). Similarly, if the tests systematically refrain from including certain content, this would create an incentive for teachers to exclude this content, in order to have more time for content that is more likely to appear on the tests. As suggested by a large teacher survey by Pedulla et al. (2003), these effects can be expected to be amplified by the severity of the stakes attached to the test results. Consequently, a test *determining* students' grades is likely to have a stronger effect on teachers' pedagogical practices, as compared to a test that is used to *moderate* teachers' assessments.

As reported by Au (2007), content alignment is a dominant theme found in research on high-stakes testing and curriculum. In a clear majority of the studies (69%), participants reported instances of the narrowing of curriculum, where non-tested subjects were excluded from curricular content. An interesting empirical example is provided by Collins, Reiss, and Stobart (2010). These researchers performed a telephone survey with 600 respondents in combination with eight focus group interviews, involving a total of 74 participants in both England and Wales. According to the teachers in this study, test preparation was perceived to narrow the science curriculum in England, since many aspects of investigatory science were reduced due to the sole reliance on paper-and-pencil items on the tests. In contrast, practical science activities, including investigations, were reported as becoming an important feature of science lessons in Wales since the abolition of national testing for 11-year-olds in science.

In Sweden, the National Agency for Education has conducted stratified sampling surveys about the steering effects of national testing. These show, for instance, that the share of teachers who perceive that their choice of curricular content is highly affected by the tests has doubled during the last decade – from one-third of the teachers to two-thirds. The latest survey also suggests that teachers with less experience are affected to a greater extent, as compared to teachers with more than 20 years of teaching experience (SNAE, 2016).

In addition to teaching experience, Lidar, Lundqvist, Ryder, and Östman (2017) suggest, based on successive interviews with teachers ($n = 16$), that science teachers may react differently to high-stakes testing, depending on whether their 'educational

philosophies' are challenged by the tests or not. For instance, one of the teachers in their study taught science primarily to get as many students as possible interested in science, which meant doing 'fun experiments', while the national curriculum and the tests prioritised planning, performing, and evaluating systematic investigations. This teacher therefore claimed that she: 'will be working a lot more with systematic investigations', 'because I want them to be successful on the national tests. But there I end up with a conflict of interest, because I do not think it is very interesting' (p. 14–15). Since this teacher adjusted her teaching to be better aligned to the national curriculum, this could be seen as a positive effect of high-stakes testing, at least from the authorities' point of view.

Consequences of high-stakes testing on teachers' assessment practices

A common assertion is that teachers, in response to high-stakes tests, narrow the focus of their assessment practices, so that students respond to only the item types found on the test, which in the US context means multiple-choice items (e.g. Madaus & Russell, 2010/2011). However, this narrowing of assessment formats is not necessarily linked to multiple-choice items. For example, Gunnemyr (2011) performed an interview study, where he compared the perceived consequences of two external tests in history: one voluntary (Swedish) and one mandatory (Finnish). Findings suggest that the Swedish teachers perceived the voluntary test to greatly influence their practice, towards increased collegial collaboration and professional conversations, while it had a very limited impact on their teaching and assessment practices. The Finnish teachers, on the other hand, perceived that the mandatory test primarily affected their teaching and assessment practices, for instance by influencing how to construct and use test items to measure students' knowledge in history. The mandatory test therefore contributed to a uniformity in terms of how to assess, although not towards an increased use of multiple-choice items, but towards an increased use of essay items supported by documents, such as historical sources or statistical information.

Examples where teachers expand their assessment practices in response to high-stakes tests are not as numerous as examples of a narrowing influence. The example provided by Lidar et al. (2017) could be a possible candidate, but it is not clear whether the teacher only teaches systematic investigations more or whether she also assesses students' skills in this area. Another possible example could be the national tests in mathematics in Sweden, which have included oral tasks since 1998. In the beginning, a majority of the teachers were either hesitant or did not think that the tests should include oral tasks, as they were time-consuming and difficult to assess. Since these first attempts, teachers' attitudes have changed (Kjellström, 2001), but it is not known to what extent teachers have incorporated a broader repertoire of assessment formats in their own teaching. Although not specifically investigating oral tasks, Boesen (2006) shows, based on a study of a stratified and random selection of teacher-made tests ($n = 52$), that teacher-made tests tend to differ from the Swedish national tests in mathematics. He therefore concludes that the influence of the national tests on teachers' development of own tests is 'fairly modest' (p. 46).

The Swedish national tests in science

The Swedish National Assessment in science will be described in some detail, since these tests may differ in several respects from other science tests around the world

(see Appendix 1 for sample items). Also, although there are minor differences between the tests for 12-year-olds and the tests for 16-year-olds, the presentation will focus on the former. The main reason is that the structure of these tests is closer to the national curriculum, which means that these tests include a broader range of different items, both regarding subject matter included and assessment formats. These tests could therefore be assumed to have a greater (positive) influence on teachers' pedagogical practice. It should be noted, however, that the tests targeting this age group and these subjects, have been provided only during three consecutive years. In the first year (2013), the tests were trialled nationwide with the whole cohort of 12-year-olds (approximately 100,000 students); during the second year (2014), all students in the country performed the tests and during the third year (2015), the tests were made voluntary for the schools. As of 2016, no more national tests in science are planned for this age group.

First of all, there are three tests, one in each subject (biology, chemistry, and physics), but each individual student only do one of them. Some weeks before the tests are scheduled, each school is randomly assigned one of the tests. Each test, in turn, consists of three parts, for which one hour of testing time is allocated and they focus on: Argumentation skills (part A), Investigations (part B), and Factual and conceptual knowledge (part C). This structure is similar for all three subjects.

Part A, which targets students' argumentation skills, consists of three constructed-response items, each focusing on a particular subskill or 'aspect of argumentation'. As an example, in the year 2013 biology test, the context in one of the items was children discussing the advantages and disadvantages of taking part in a soccer competition if you have a cold and are not feeling well. In the task, three fictional characters give their opinions in speech bubbles and the students are expected to suggest how to continue the conversation.

Part B targets students' skills in planning, carrying out, and evaluating systematic investigations. As an example, in the year 2013 biology test, there were six constructed-response items. One was about asking questions possible to investigate, two about planning an investigation, two about carrying out small investigations, and one about evaluating an investigation.

While parts A and B address fairly well-defined subskills, Part C aims to assess a broad range of factual and conceptual knowledge, such as identifying predatory animals in the biology test or describing the properties of magnets in the physics test. Furthermore, while almost all items in parts A and B are constructed-response items, the items in part C are a mix of both selected- and constructed-response items. Since all parts of tests are allocated one hour of testing time, there are roughly six times as many items in part C as compared to part A.

A particular feature of the Swedish National Assessment tests is that they are scored by the teachers themselves. To that end, each test is accompanied by a comprehensive set of instructions for how to score each of the items in the tests. There are also instructions about how to summarise the item scores from all parts of the tests (A, B, and C), in order to generate a test score (or more truthfully a 'test grade' from F to A, where F is fail, E is pass, and A is the highest passing grade) for the test as a whole.

Finally, it should be recognised that the tests are developed according to a rigorous methodology, involving peer review by researchers in the field of science education, review panels of in-service teachers, and trials with hundreds of students from different

geographical regions. Furthermore, both quantitative and qualitative analyses of item and test data are performed as part of the test development, before as well as after the tests have been administrated, including differential-item-functioning analyses and estimations of interrater agreement.

Summary and research aim

Tests convey messages about both *what to teach* and *how to assess* student performance in relation to this particular content. As has been shown, both of these dimensions may either broaden or become more uniform and narrow as a consequence of high-stakes testing. Current research suggests that the most widespread consequence is the narrowing of the curriculum, where teachers pay less attention to (or even exclude) subject matter that is not tested. Which the consequences are depends on the interaction between test design (e.g. subject matter included and assessment format) and teachers' teaching and assessment practices, but also factors such as teaching experience and 'educational philosophy'.

The specific context addressed in this article is the influence of the Swedish national tests in science. These tests are of particular interest since they include non-traditional content, such as argumentation skills, which is not necessarily part of teachers' teaching repertoire, as well as content that may be present in traditional science teaching, but not always assessed (such as systematic investigations). These tests also include a variety of different assessment formats, such as simulated conversations, writing letters, as well as planning and carrying out investigations. These tests could therefore be assumed to contribute to an expansion of the curriculum, as well as a broadening of different assessment formats. In is currently not known, however, neither how, nor to what extent, teachers' pedagogical practices are influenced by these tests.

As described above, surveys performed by the Swedish National Agency for Education (2016) suggest that Swedish teachers' choice of curricular content is highly affected by the tests. These surveys do not, however, provide any specific information about science teachers or present any details of how the curricular content is affected (for instance, whether it is broadened or narrowed). This study therefore aims to investigate how Swedish science teachers are influenced by national, high-stakes testing in science, specifically focusing on instances where teachers' pedagogical practices are broadened and/or narrowed.

Methodology

The overall design of this study is qualitative thematic analysis of focus group data from group discussions with Swedish science teachers. The data are longitudinal and originates from a larger study, investigating teachers' perspectives on the role of 'nature of science' (NOS) in science teaching (Lidar 2, 2017) during three consecutive years. In this study, the data from the focus-group discussions have been re-analysed from a different theoretical perspective: the influence of high-stakes testing in science. This analysis is made possible by the fact that the participating teachers continuously discussed the tests during the meetings, even though this was not a topic initiated by the researcher. The data for this study have therefore been produced unintentionally and coincidentally, which has both advantages and disadvantages for the current research.

A major advantage, which is also the main argument for using this data, is that the frequent and unprovoked discussions about the national tests indicate how important this topic is for these teachers. As soon as they come together, and as soon as an opportunity appears, they start discussing the tests, although they meet for a different reason and are expected to discuss other topics. This suggests that the tests are more or less omnipresent in the minds of these teachers and therefore a highly relevant subject for investigation. Another advantage is that the discussions about the national tests are unaffected by the interviewer/researcher, since the sanctioned attention in the focus groups is on different perspectives on teaching NOS.

One of the main downsides of using data from another project is that it might be less focused, coherent, and complete. These disadvantages are obvious in this case, where the discussions about national testing are scattered across the data and the discussions are sometimes brief, making the analysis problematic. On the other hand, since the discussions are numerous and the fact that some of them were already transcribed, facilitated the analysis. In order to make the material more complete, data not used for the previous project (such as dialogue before the actual focus group discussions) has been included in the current data set.

The sample

The sample in this study consists of one group of teachers ($n = 7$), who were involved in the project mentioned above. The group was comprised of teachers from four different schools, although half of the teachers belonged to the same school. All but one of the teachers had long teaching experience (i.e. 10 years or more; mean = 16 years). One of the teachers in the group was no longer teaching science after two years in the project and left the group. She was replaced by another teacher at the same school from the beginning of year three. All but one of the participating teachers were female.

There were some differences in the educational backgrounds among the participating teachers. All participants had received a teacher education, but the amount of science courses within their respective teacher-education programs differed. One of the older teachers had a general teacher education for all subjects taught in years 4–6 in Swedish compulsory school. The younger teachers had teacher educations with a mathematics and science profile aimed at years 1–7 or years 4–9. The teachers also taught either primary, or lower secondary, science, which means that they had experience of either the national tests for 12-year-olds or the tests for 16-year-olds. The characteristics of the sample are summarised in Table 1.

Table 1. Overview of the sample.

Teacher	School	Teaching experience	Sex	Grades
1	A	15 years	Male	6–9
2	A	10 years	Female	6–9
3	B	32 years	Female	4–6
4	A	14 years	Female	6–9
5	A	8 years	Female	6–9
6	C	17 years	Female	4–6
7	D	16 years	Female	1–6

The focus groups

In the original project, focus-group discussions (Morgan, 1997) were chosen for the purpose of gaining insight into teachers' perspectives and transforming perspectives, as they gained experience from taking part in activities and discussions in the focus groups. The methodology resembled participatory research (Wibeck, 2010), since there was not just one focus-group discussion, but several discussions taking place during a longer period of time. The purpose of such discussions is not only that the researcher may gain insights into participants' views and experiences on a certain topic, but also for initiating and sustaining a developmental process among the participants.

The focus-group discussions (12 in total) were more or less evenly distributed over three years. Each discussion lasted for two hours and they were mainly carried out during the participants' working hours. In connection to the discussions (either before or during) participants were often provided with some working material, in order to stimulate the dialogue. The material, which consisted of short texts to read beforehand or tasks to work with during the meeting, focused on perspectives on teaching NOS (i.e. not assessment or testing).

All focus-group discussions were recorded and transcribed. In the first transcripts, everything was transcribed verbatim, even peripheral small talk. However, in the later transcripts parts of the conversation that clearly did not deal with the topic of the study (such as teachers' talking about national testing) were left out of the transcript, but with a reference to what the conversation was about. Data for the current study are therefore partly transcripts and partly audio data from the focus-group discussions.

During the focus-group discussions, the researcher took responsibility for making room for all participants' voices. Some of the participants in the group knew each other well and had been working together for a long time, while others had never met prior to the first meeting. All of the participants were experienced teachers, however, and were willing to share ideas and did not hesitate to question each other.

It is unavoidable that the researcher influences the actions of the group, for instance by directing the group's attention towards certain topics. In this case, however, since the original focus was on the teaching of NOS, the researcher cannot be assumed to have substantially influenced the discussion about national testing.

The analysis

The interview data were analysed with conventional thematic analysis, which is a method for identifying, analysing, and interpreting patterns of meaning (or 'themes') within qualitative data (Clarke & Braun, 2017). The analysis followed the procedure outlined by Braun and Clarke (2006), which, in this case, means that the following steps were taken:

1. The first step was to listen to the recordings and create time logs in spreadsheets for those parts of the data that were not transcribed, so that all data could be searched and organised.
2. Interesting features of the data were coded across the data set and the data were organised in relation to statements about the influence of national testing. The coding was performed by one of the researchers and then checked by the other. Any disagreements were discussed and resolved by consensus decisions.

3. Codes were assembled into initial themes in relation to how teachers' pedagogical practices were influenced by the tests, gathering data relevant to each initial theme. Initial themes were discussed among the researchers and then refined.
4. Revised themes were checked against coded extracts and the data set as a whole.
5. The specifics of each theme were refined.
6. A selection of compelling extract examples for this article was made, in order to substantiate the proposed themes. The selection was also made to show a breadth of the data and to represent the different voices of the teachers in the sample.
7. The extracts were translated to English by the authors.

Findings

Discussions pertaining to national testing in science occurred in 11 of the 12 focus-group discussions analysed, although the extent varied greatly. In one end of the spectrum, national testing was only mentioned once by one teacher (focus-group meeting 3, year 1), while at the other end, discussions about national testing made up a significant part of the meeting (focus-group meetings 2 and 3, year 2). In most cases, however, national testing was discussed at one or two occasions during each meeting. It should also be noted that at three occasions, teachers spontaneously started to discuss national testing when they met, before the actual meeting. These discussions are also part of the empirical material in this study.

Below, the findings are presented as transcending themes found in the data, representing ways in which the teachers are influenced by the national tests in science. The six teachers are designated T1–T6 in relation to the quotes, and the interviewer as 'I'.

The tests are used as a proxy for the national curriculum

Although the tests, as well as teachers' teaching and assessment practices, should ideally be grounded in the national curriculum, a recurring theme in the discussions is that the teachers use the tests as a substitute for the national curriculum. According to the teachers, the tests let them know how the Swedish National Agency of Education (i.e. the mandating authority of the tests, as well as the authority responsible for the national curriculum) thinks, and the more tests you get to see, the better your chances are to decipher the implicit messages conveyed by these tests:

> / … / the more examples that you see, the more national test you see, the better feeling you get about how they think. (T1)

For the teachers, what is written in the national curriculum is of minor importance, since it is the operationalisations made in the tests that have an impact and that provide direction:

> I Is this a scientific argument? / … / It's in the curriculum, to be able to distinguish science from other perspectives, but I'm not … I mean, I haven't looked at the tests.
> T2 No, it's in the aim and in the core content, but if I remember correctly, there's nothing about this in the standards.
> T1 In the physics test, the national standards, I don't know, but it hasn't been on the *physics* tests anyway. But it could have been in another part.

T2 It says 'separating values from facts', that's what it says.
T1 Then it's there.
T2 But I don't think it differs, I mean, distinguishing scientific … I don't think it's explicit in the standards, but it's there, but the real question is how it will be manifested in the national tests …
I You've talked about the national tests during these sessions as well, how they are designed, that they …
T1 … and then you kind of get influenced by them.
T2 I think it's a combination …
T1 Quite a lot of the national tests influence how we teach, because it becomes a kind of practical interpretation of …
T2 Where we're going …
T1 … the national curriculum. This is kind of how it should be done.
T2 [Agrees]

The main reason for teachers to consciously let themselves be influenced by the tests is that they do not want their students to fail the tests – or as expressed by one of the teachers (T3), it would be 'bad luck' for the students if they were to be tested on something that they have not been taught. This also means that the students need to practice how to solve the kind of items found in the tests, so that they understand 'how it works':

T1 No, I'm thinking that they [the students] should read through this … They've never solved a similar task before, so it'll be some practice and I'll assess it formatively. Really, I should have made them solve the one on the national test last spring, it's been made public …
I However, it's not the same aspects … [that are being assessed]
T1 No, but they'll understand how it works.
T5 We've previously discussed to do a small assignment that's very restricted [similar to a test item], just in order to, kind of, what it is, what it is that the teachers [and the test developers] want, because we have some of them who … [have not understood how it works]
T1: We've started preparing by doing last year's tests, I mean using them as benchmarks and we say that we're trying to build a bit on that and …

The inclusion of new aspects of school science

As described above, the Swedish national tests in science for 12 year-olds were composed of three parts, focusing on Argumentation skills (part A), Investigations (part B), and Factual and conceptual knowledge (part C), reflecting a tripartite conception of school science as outlined in the national curriculum. The same three aspects (A–C) can be found in the tests for 16 year-olds as well, but not necessarily organised as three equal-sized parts. Rather, students are usually only given one (larger) task about planning and carrying out a systematic investigation, while the 12 year-olds were given several (smaller) tasks focusing specifically on posing questions, formulating hypotheses, planning, carrying out, or evaluating investigations.

In the discussions among the teachers, it is clear that neither argumentation skills nor planning and evaluating systematic investigations have been part of their previous pedagogical practice. Since these aspects are now part of the national tests, however, the

teachers have felt forced (although not necessarily in a bad way) to include them in their teaching.

T3 / … / in the national tests there are always some [items about] planning … or improving [investigations]. It is very clear that these are the things to work with …
T1 Yes, sure.
T3 It is quite interesting to use these [items]. I have given them to year four [students] now that we finished [working with] energy and physics, because they were supposed to plan an investigation and to be able to carry out [the investigation] and there were year four [students] who could handle it very well. It was this [item] with nails, checking whether the salt made it [the nail] rustier than with just water, or not at all or how it worked. They managed to explain it very well, as compared to some other students, who were older. So it was a bit funny to see …
T2 We're laughing now, because we do this with [students in] year eight and nine.

It is also evident from the discussions that the teachers perceive the inclusion of planning and evaluating systematic investigations as a legitimate extension of the enacted curriculum:

T1 … when you have given them [the students] the opportunity to carry out an investigation, you try to let them think about whether this investigation raises any thoughts about other investigations, I mean this aspect about developing it further … as they have included in the national tests, where they [the students] consider the quality of the investigation, as well as how to refine the investigation, try to improve their investigations. Not only [think] 'OK, now we've done this and we're finished', but 'Next time we're doing this again, only better, so what are we going to change?', and 'Why [do you think it] will it be better if you do like this?'.
T6 If you think about the national tests, there it's 'use your knowledge in science' or 'which questions can you get good answers to, or perform your own investigation about'. 'Write your own questions', I mean, they [such assignments] aren't easy to find, or to come up with such tasks by yourself just like that, and start that way of thinking [among the students].
T7 But at the same time, there wasn't any [such item] in the test, or there was when we worked with it. They [the students] had to take a stand about what they thought, and then they had to use their knowledge in science to argue. It wasn't the test, it was when we had our … when we read about different energy sources.
T6 I think there was. Yes, no, it's correct. I think there was one [item] about wind turbines, where you were supposed to argue …

The curricular extension in relation to planning and evaluating systematic investigations also involves a negotiation/broadening of what it means to do science, for instance regarding the role of imagination and creativity:

T6 I think it becomes rather clear. We did, we've done the national test in physics, so I tried the examples made available on the National Education Agency website, where they were supposed to prepare for investigations, as you said, but now it was biology and with a plant; can a dandelion survive in total darkness?
T1 Mmm …
T6 Prepare an experiment or an investigation about that; first, they were sitting like this, right? And then, when they started, they had to think about it and then discuss in small groups, then discuss in larger groups, so they realized that they had thought about different things and that you need to be a bit creative and not only do it in one way, but that you have to think about several aspects. So I think that although they do not say 'I'm imaginative, so I can do this' …

T1 [They] do not articulate it.
T6 they still notice it.

A limited interpretation of argumentation skills

In contrast to planning and evaluating systematic investigations, argumentation skills are *not* seen as a legitimate extension of the enacted curriculum, at least not the science curriculum. Rather, the teachers perceive the items addressing these skills as too extensive and time-consuming/difficult to assess, and as a part of language learning instead of science:

T2 I've not been thinking too much about these things before, because this has kind of not been part of science teaching for me, not the traditional science teaching.
I Do you think is a part now?
T2 Yes, to a greater extent, I think.
T1 Yes, these things, yes. The essays that you're supposed to assess. Makes me angry just thinking about it.
I Which essays?
/ … /
T1 This part of the national tests, about argumentation, I think it is about language learning. Identifying chains of reasoning and counting layers and …
T4 I think they are too extensive as well, because you have to make comparisons with everyone and with everything … Well, I'm not really sure …
T1 I think it is too much of language learning …
T4 Yeah, well, it's too extensive …
T2 I think so too.

Since the teachers are reluctant to include argumentation skills in their teaching, but still do not want their students to be at a disadvantage when doing the tests, the teachers search for strategies to handle this dilemma. The main strategy is to include only enough for the students to pass the test. In the quote below, the teachers have noted that the tests do not contain authentic texts for the students to read, when preparing for writing argumentative tasks. Instead, such test items may provide information in a more digested format, such as post-it notes or fact sheets:

T1 No, I was thinking that … and then we have used a number of texts, but just reading those articles and being able to remember and notice things. However, in the national tests they have a fact sheet that you can read three-four times.
T2 Where everything is brought together.
T1 I mean, it is compiled. In that case, we should have worked with these texts more from the beginning, maybe in cooperation with the [teacher in] Swedish by trying to produce such a fact sheet ourselves from the texts.

In the next quote, the teachers suggest that the students should write their answers as bullet points, rather than writing coherent texts, as is done in one of the tests:

T1 … it is these three tasks, which really makes it much more complicated.
[Everyone agrees]
T1 … and much more of a language-learning task. However, you can write it as bullet points, because then it is less language learning. Then you've practiced thinking about it and it's nothing wrong with that, but it's just that we wanted them to use this for writing and using it for argumentation, and they haven't really managed

that, but maybe it's too difficult for them. I don't know what you can expect from them, really.

/ … /

I Do you think that's language learning too?

T1 No, if they [write their] answer as bullet points, as the students have done to a large extent, then maybe it's more of a science task. It's when you start using it in a text and for argumentation …

When formulating the assessment criteria, the teachers also seek guidance in relation to the tests:

I What kind of rubric are you working with here?

T1 I'm thinking about the one used in the national test.

T2 Provide one argument, provide two arguments, provide three arguments.

T1 Different aspects.

I Different arguments with different aspects then?

T2 [Agrees]

T1 It's not really clear exactly what [difficulty] level it should be and the only thing with levels that we can start from, is by looking at the requirements in the national tests and it feels like it provides a level of difficulty in some way.

Discussion

The aim of this study was to investigate how Swedish science teachers are influenced by national, high-stakes testing in science, which was done by a thematic analysis of focus-group discussions.

Previous research has suggested that tests convey (explicit and/or implicit) messages about both *what to teach* and *how to assess*, and that both of these dimensions of teachers' pedagogical practice may either broaden or become more uniform and narrow as a consequence of high-stakes testing. However, the most widespread consequence seems to be a narrowing of the curriculum, where teachers pay less attention to (or even exclude) subject matter that is not tested (e.g. Au, 2007; Gunnemyr, 2011; Minarechová, 2012).

The findings from this study corroborate these previous findings, by showing that when new content is introduced by the tests, the teachers respond by including these new aspects of school science into their pedagogical practices (i.e. a broadening of their pedagogical practice). This is done quite consciously by the teachers, since they openly use the tests as a proxy for the national curriculum and, naturally, do not want their students to fail the tests. However, the findings from this study also suggest that the reception of new content may differ, depending on how legitimate the teachers perceive that the new content is (cf. Lidar et al., 2017). If the new content is close to their current teaching traditions, the teachers may incorporate the new content into their existing practices. This was clearly the case with planning and evaluating systematic investigations, but not with communication skills.

On the one hand, this could seem reasonable, since systematic investigations are often seen as a part of 'traditional science teaching', while a focus on communication skills would rather be categorised as progressive or critical school science, if using the terminology from Zacharia and Barton (2004). This is particularly true if these skills are linked to socio-scientific issues. On the other hand, lab-work in traditional school science is often guided by a strict manual, and aims at establishing students' conceptual understanding

and/or mimic the work of real-life scientists, while the test items (in this case) require students to pose questions and design their own investigations. The national tests may therefore be argued to support a traditional view on science teaching, as well as a more progressive/critical view, depending on the context in which the investigations take place. If placed in a science context, the investigations are closer to a traditional view, but if placed in a broader, societal context, they are closer to a progressive/critical view. As the actual test items can be placed at different points along this gradient, this means that teachers can choose to implement an interpretation of planning and evaluating systematic investigations, which is in line with their own teaching traditions.

In theory, teachers should be able to handle the implementation of communication skills in a similar manner, since communication skills could also be placed in a science context, for instance by focusing on how students support their arguments with data from scientific investigations, or by mimicking the work of real-life scientists when communicating findings to peers. In this case, however, the test items are almost exclusively situated in a broader, societal context, which possibly makes it more difficult for teachers to translate the new content, so that it may fit into a traditional view on school science. The teachers are therefore seen to stay close to the interpretation and operationalisations of the national curriculum made by the test developers, by using the same criteria and similar item formats, without taking into consideration that these items and criteria are specifically designed to meet challenges of standardisation and high interrater agreement in scoring. This means that although the tests contribute to a broadening of teachers' pedagogical practices, by implementing communication skills into their teaching and assessment, the tests at the same time narrow teachers' interpretation of how communication skills can be taught and assessed. For instance, teachers are seen to refrain from using authentic sources of information, instead wanting to use digested formats, such as pre-prepared fact sheets. While using pre-prepared fact sheets can be judged an adequate strategy to meet the needs of standardisation, where all students must be able to read the same material during a relatively short period of time, similar restrictions do not necessarily apply during regular teaching. Similarly, while using the number of arguments that students formulate as an assessment criterion, rather than the quality of the arguments, can be judged an adequate strategy to meet the needs for reliable scoring on standardised tests, similar restrictions do not necessarily apply during classroom assessment. In addition, according to the national curriculum, students are expected to learn how to search for information, using different sources, and reason about the usefulness of both the information and the sources. Searching for information is, however, difficult to test with a paper-and-pencil test, and is therefore not included in the national tests. This means that this is an aspect that runs the risk of being excluded from the teaching as well.

Conclusions

The findings from this study suggest that the Swedish national tests in science may influence teachers' pedagogical practice by being used as a substitute for the national curriculum. Since the teachers do not want their students to fail the tests, they consciously implement new content that is introduced by the tests. However, in cases where this new content is not seen as a legitimate part of teachers' established teaching traditions, the interpretation and implementation of this content may more or less replicate the

operationalisations made by the test developers, even though these operationalisations are severely restricted by demands for standardisation and reliable scoring. Consequently, the tests simultaneously broaden and narrow teachers' pedagogical practices.

Limitations and suggestions for future research

This is a small-scale study, including only a few teachers (although they participated during an extended period of time), which means that the views expressed by the teachers in the sample are not necessarily representative for other teachers in Sweden or elsewhere. It should also be noted that data from focus-group discussions represent the perceptions and beliefs of the participating teachers, not necessarily their actual practice, although in this case the discussions are authentic in the sense that they were not initiated or moderated by the researchers.

The findings are based on a qualitative, thematic analysis of focus-group data, which means that the main contribution of this research does not lie in generalisable findings, but in a deeper understanding of how high-stakes testing may influence teachers' pedagogical practices. In this case, this understanding is linked to the Swedish national tests in science, which are, in some senses, unique, since they include a broad repertoire of different item formats and also target different aspects of science teaching, including communication skills and planning systematic investigations. As these tests could be expected to have a positive influence on teachers' pedagogical practice, for instance by including new content and broadening their use of different assessment formats, the findings from this study suggest that this influence is not necessarily unambiguously positive. Rather, the tests may contribute to a narrow and limited interpretation of how to teach and assess communication skills in science.

It should be noted, however, that most of the data supporting a limited interpretation of how to teach and assess communication skills in science comes from the teachers teaching years 7–9 (i.e. lower secondary school), while less is said by the primary teachers. This *could* indicate that these teachers are more comfortable with teaching and assessing these skills, for instance since some teach first and foreign languages and social sciences, as well as science, but it is not possible to make any such claims based on the current data.

Since it is not known to what extent the findings are generalisable to a larger population of science teachers, a natural recommendation for future research could be to include a larger sample of teachers. This could be done, for instance, in a survey asking science teachers in more detail (as compared to the general surveys by the Swedish National Agency of Education), about how they perceive that their teaching and assessment practices are affected by national tests. Still, as teachers may not necessarily be aware of the consequences from the tests or, as in this case, where the influence is multifaceted, it may be a challenge to find adequate items for such a survey. Perhaps it would therefore be more appropriate to first look more closely at teachers' teaching and assessment practices. In this study, teachers discuss their teaching and assessment in relation to the national tests, but the discussions did not specifically address the influence from national tests, and the tasks or the criteria that the teachers discussed are not part of the data. It would therefore be interesting to collect teachers' actual teaching materials and ask them to present and justify this material, as well as to compare it to the design of the national tests (cf. Boesen, 2006). In such a study, it could be possible to analyse more closely the specific influence of the tests on teachers' pedagogical practice.

Notes

1. The term 'pedagogical practices' refers to both teaching and assessment practices, as well as both *what* to teach/assess and *how* to teach/assess. In this context, however, the term is limited to *what* teachers teach about (content) and *how* they assess (assessment format), reflecting the empirical focus of previous research on the impact of external testing.
2. This model also makes a distinction between the test developers and the mandating authority, such as a state department of education, who can communicate independently with the schools.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Anders Jonsson http://orcid.org/0000-0002-3251-6082
Lotta Leden http://orcid.org/0000-0002-8255-3607

References

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*, 258–267.
Author 2. (2017).
Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa*, *80*, 139–148.
Boesen, J. (2006). *Assessing mathematical creativity (Doctoral dissertation)*. Umeå: Umeå University.
Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101.
Clarke, V., & Braun, V. (2017). Thematic analysis. *Journal of Positive Psychology*, *12*, 297–298.
Collins, S., Reiss, M., & Stobart, G. (2010). What happens when high-stakes testing stops? Teachers' perceptions of the impact of compulsory national testing in science of 11-year-olds in England and its abolition in Wales. *Assessment in Education: Principles, Policy & Practice*, *17*, 273–286.
Gunnemyr, P. (2011). *Likvärdighet till priset av likformighet? [Equivalence at the cost of uniformity?] (Licentiate thesis)*. Lund: Lund University.
Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice*, *10*, 169–207.
Kjellström, K. (2001). Muntlig kommunikation i ett nationellt prov [Oral communication in a national test]. *Nämnaren*, *2001:2*, 41–47.
Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy & Practice*, *22*, 302–323.
Kohn, A. (2000). The case against standardized testing: raising the scores, ruining the schools. Retrieved from: http://teacherrenewal.wiki.westga.edu/file/view/Testing,+Testing,+Testing.pdf
Lidar, M., Lundqvist, E., Ryder, J., & Östman, L. (2017). The transformation of teaching habits in relation to the introduction of grading and national testing in science education in Sweden. *Research in Science Education*. doi:10.1007/s11165-017-9684-5
Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh yearbook of the national society for the study of education* (pp. 83–121). Chicago: University of Chicago Press.
Madaus, G., & Russell, M. (2010/2011). Paradoxes of high-stakes testing. *Journal of Education*, *190*, 21–30.
Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy*, *3*, 82–100.

Morgan, D. L. (1997). *Focus groups as qualitative research*. Thousand Oaks, CA: SAGE.
Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy.
Swedish National Agency for Education. (2016). *Nationella proven i grundskolans årskurs 6 och 9* [The national tests in year 6 and 9 in compulsory school] (Report no. 447). Stockholm: Swedish National Agency for Education.
Swedish National Agency for Education. (2018). *Curriculum for the compulsory school, preschool class and school-age educare*. Stockholm: Swedish National Agency for Education.
Wibeck, V. (2010). *Fokusgrupper: om fokuserade gruppintervjuer som undersökningsmetod [Focus groups: On focused group interviews as a research method]*. Lund: Studentlitteratur.
Zacharia, Z., & Barton, A. C. (2004). Urban middle-school students' attitudes toward a defined science. *Science Education*, *88*(2), 197–222.