Uppsala University
Department of Education Resarch report, April 2019

Analysing validity: The case of Swedish national tests in year 6 science

Jonas Almqvist, Graham Orpwood, Eva Lundqvist & Malena Lidar

The purpose of this article is to analyse and discuss standardized tests in biology, physics and chemistry with a special focus on their content validity. In the article we describe and discuss three different tensions between the Swedish curricula and standardized tests in science: (1) Curricular intentions and assessment choices, (2) The 'knowledge requirements' specified in the curriculum and the marking scheme used in the assessment and (3) The intention of the evaluation system and its actual result. These tensions have consequences for the validity of the tests. Hence, it is necessary to regard these tests as only one of many resources teachers can use in their teaching and assessment practices.

Keywords: validity; standardized tests; science education

Introduction

During the period 2013-2015 national tests were given in each of biology, chemistry and physics in year 6 in Swedish compulsory school. These tests were part of an ongoing series of reforms in the Swedish educational system, including a new science curriculum and a new grading system. After a couple of years, the national tests for year 6 science were taken away due to the heavy work load on the teachers. However, from March 2017 all the tests given during the period 2013-2015 are available for teachers in a database run by the Swedish National Agency for Education as examples for teachers to use in their planning of teaching and assessment.[1] Thus, even though the tests are not mandatory anymore, they may be regarded as the governments' interpretation of the curricula and how students' knowledge should be assessed. Since the tests are made available through the database and may be

[1] https://bp.skolverket.se/, visited October 17, 2017.

regarded as exemplary and as standardized tests, it is also important to study them and what they may imply for teaching and assessment of the students knowledge in science.

The major aims of the national tests in year 6 science education were to support equal and fair assessment and grading of the students' knowledge, and to provide a basis for analysing the extent to which the knowledge expectations of the curriculum had been met. A third aim, to support teachers' work by concretising the curriculum and subject plans, which along with the two other and an additional aim about increasing pupils' target fulfilment, was the original ambition of the reform.

After the decision to end the national tests in science and social sciences at year 6 only the third aim, to support teachers' work, remains for these subjects. In 2015, the Swedish National Agency for Education in its information about the national tests, wrote that

> *If you compile the class results from the national test, you can see how well the class as a whole performed. Such a compilation can help you as a teacher to evaluate your teaching and give signals about the areas that may need to be strengthened in the future (Skolverket 2015a, p. 14).*

From March 2016, the Swedish National Agency for Education build a database with assessment tasks similar to the questions given in the national tests in order to provide teachers with good examples of assessments.

> *We will provide assessment support in the form of tests in the Assessment Portal. One earlier national test in each subject will be published as assessment support from March 2016. At the same time, work is underway to develop new assessment support that will be published in the Assessment Portal in early 2017.*

From March 2017 all national tests given are available through this databes. In addition, a few more examples are uploaded. Hence, a central part of the new assessment support that the Swedish National Agency for Education write about in the quote above consists of the national tests from 2013-2015. In this article we will analyse the national tests uploaded in the assessment portal.

During the period when the tests were given, schools were randomly assigned to do one of the tests in biology, chemistry or physics. The content of the national tests was based on national policy documents, and, in particular, the subject plan contained in the curriculum. The tests

were formulated centrally, and distributed to schools, but the teachers themselves were responsible for marking them, in accordance with a given marking scheme.

The reaction to these tests from teachers has been mixed; some teachers were clearly negative, but many also expressed the view that, even though the tests demanded a lot of their time, they provided them with perspectives both on teaching content and on student knowledge which they would not have otherwise had (Hirsch, 2016; Lidar et al., in review; Ryder et al., in review). However, the argument that too much administrative time had to be spent on the tests, which is described in the National Agency of Education's own evaluation of the reforms (Skolverket 2015b), clearly resulted in the decision to remove them. Even though the national tests in year 6 science are no longer mandatory and no can longer be considered as high-stake tests, they are still placed in an assessment portal an are thereby a potential support for teaching. If these tests are supposed to provide support for the teachers and concretise the curriculum and subject plan, we need to ask if they actually measure what they are expected to do, or if there are any significant tensions between what the students are supposed to know and what the tests measure? The purpose of this article is to analyse and discuss the national tests in biology, physics and chemistry with a special focus on their content validity.

Standardized tests in science education

The relation between the use of standardized tests, teachers' teaching and the changes of educational practices is complex and depends on teachers' knowledge, approaches to teaching and views on learning (Cimbricz, 2002). Nevertheless, studies have shown that the use of standardized tests often influence teachers' selection of teaching content and are often narrowing it to what is expected to be tested (Au, 2007). In Sweden, the national tests, even though they were given during such a short period of time, they have influenced many teachers' way of thinking about assessment in science education (Hirsh, 2016; Lidar et al., in review; Ryder et al., in review).

One of the aims of the tests was that they are supposed to provide the teachers with a tool for interpreting the science curriculum and providing them with examples of how the students' knowledge may be assessed. However, it should also be noted that, even though it may be reasonable to implement standardized tests in order to change teaching, the tests may not be used by teachers in any conform way (lidar et al., in review). This is probably even more so now the tests no longer are mandatory. Black, Harrison, Hodgen, Marshall and Serret (2010)

show in an intervention study of how to improve teachers' use of summative assessment, that the teachers' understanding of validity needed to be challenged in order to use the assessment for development of teaching. They also worked together with teachers to develop teaching and were engaged in discussions about validity. The study show, among other things, that there may be a tension between teachers' way of assessing on one hand and their belief about validity on the other.

Furthermore, the goals and selection of educational content in science teaching, and hence what should be tested, has changed historically and may differ between countries (cf. Fensham, 1988; Lidar, Karlberg, Almqvist, Östman, & Lundqvist, 2017; Linder, Östman, Roberts, Wickman, Erickson & MacKinnon, 2010; Marty, Almqvist & Venturini, 2017; Roberts, 1982; Östman, 1998). One of the key findings in this line of research is that there are different parallel teaching traditions expressed in curricula, text books, teaching and in tests (cf. Almqvist & Lundqvist, 2013; Englund; 1986; Lundqvist, Almqvist & Östman, 2012; Lundqvist & Sund 2016; Östman 1996). For example, in a review of international research on scientific literacy, Douglas Roberts (2007) shows that there is a strong polarization between two different visions for science education. In principle the two orientations privilege in different ways which content that should be taught: 1) The students should learn to use basic scientific concepts, models, theories and skills in discussions and investigations of the world; 2) The students should begin the work with complex authentic problems where scientific knowledge is used. These two visions of scientific literacy – Roberts (2007) calls them simply Vision 1 and 2 – can be used to describe and discuss the differences that exist with regard to the content of science education.

Validity in tests

Given the notion that goals and content of teaching may vary, so must the discussion of the validity of tests do.[2] There should for example be a difference between a test designed for teaching in accordance with what should be privileged within vision I and vision II respectively. The question about the aims of educational assessment is a central quality question and in fact, for the last decades attention has been paid not only to the quality of the

[2] Newton & Baird (2016, p. 173) stress that "/.../there is no widespread professional consensus concerning the best way to use the term [validity]". This is not the place to dig deeper into the debate about validity, but rather focus on questions about the what we chose to call content validity of the tests analysed. More specifically, we are interested in questions about how well the content assessed in the tests corresponds to the overall aims of them.

methodologies for measuring students' knowledge but also a discussion of how the assessment tools fit the purposes of assessment (Broadfoot & Black, 2004). The issue at stake is what it means to be good at science and how teachers may assess students' knowledge in relation to this. Newton (2007) describes and discusses different purposes of educational assessment and argues that the discussion about how to use a judgment often is rather confused. He shows how different categories of purpose are relevant for different stakeholders in educational systems and stress that this should be taken into account and influence the design of tests and assessment systems.

One of the issues in discussions about educational assessment concerns the validity of performances assessment (Messick, 1994). Osborn (2013) argues that the 21st century science education and its focus on learning scientific reasoning, in contrast with recalling of facts, creates challenges not only for teaching in a narrow sense but also for how to select which competences the students should learn in science education and how to assess their performances in school. In a review of research, Sampson and Clark (2008), show how different analytic frameworks focus on different aspects and processes in scientific discussions and argumentation. The nature of and quality of scientific arguments differs and need to be assessed in various ways in order to be valid within a specific framework.

Consequently, planning, performing and assessing science education is a matter of selecting goals, content and methodology. When a standardized test is made, it cannot assess every possible aspect of teaching. The selection of content in the tests construes the kind of scientific literacy that will be measured (Almqvist & Lundqvist, 2013; Bybee, McCrae & Laurie, 2009). Hence, the quality of the assessment, and especially the validity of a test, is not only a matter of methodological issues, but also of selecting content.

In this article we take this into account, and we are specifically interested in the construction of the tests in relation to the curriculum. Orpwood's study (2001) shows how the changes in curriculum are seldom linked to changes in assessment. The results show how science curriculum reforms where a new kind of content (in his case Science, Technology and Society) was implemented in the curriculum, but also how the practices of assessment changed much later.

Content and design of the year 6 National Tests in Science

The Swedish national tests in Science were based on statements of aims and content in the subject plans in the curriculum for years 4-6 in each of the subjects Biology, Physics and Chemistry (Lgr 11, pp. 111-158). The analysis of the national tests reported in this article is focused mainly on the curriculum aims, which are set out, for each science subject corresponding to three general purposes of science education, which includes to develop: a) the ability to participate meaningfully in discussions of ways in which science, technology and society interrelate, b) the ability to conduct scientific investigations and c) a good understanding of scientific concepts, models, and theories.

As the curriculum for all three science subjects are structured in a similar way, the design and content in the physics curriculum will here be presented as example. The Swedish physics curriculum is structured in 4 parts: aim (for all years 1-9); core content (subdivided into years 1-3, 4-6, and 7-9); and knowledge requirements (for each grade level at the end of each three-year period). The aims of the curriculum are in sum to enable pupils to:

A. use knowledge of physics to examine information, communicate and take a view on questions concerning energy, technology, the environment and society;
B. carry out systematic studies in physics;
C. use concepts of physics, its models and theories to describe and explain physics relationships in nature and society.

The tests follow the structure of these three aims and as will be evident our analysis traces the relationships among the aims in the curriculum, the tasks and questions students are given to undertake, and the grading schemes to be used by teachers in evaluating students' work.

In passing we would note that these curriculum aims follow a general international trend in science education towards combining knowledge of science (aim 3), ability to conduct scientific investigation (aim 2), and understanding of the inter-relationships among science, technology, society and the environment (STSE) (aim 1). What is unusual and interesting here, however, is that the STSE aim – far from being the afterthought it is in many national curriculum statements – comes in first (and by implication the most important) place (cf. Marty et al 2017).

The core content for the curriculum is divided into three stages. The core content in years 1-3 focuses on what are described as 'science studies' – seasons of the year in nature, body and health, force and motion, materials and substances in our surroundings, narratives about

nature and science, and methods and ways of working. Years 4-6 sees physics as a discipline more prominent with the core content topics summarised as follows:

- *Physics in nature and society (energy, weather)*
- *Physics and everyday life (energy, electricity, magnetism, forces and motion, sound, light)*
- *Physics and world views (physics discoveries and their importance, different cultures and their ways of explaining, the solar system, space and satellites, measuring time)*
- *Physics, its methods and ways of working (planning and executing simple studies, measurement and instrumentation, documentation and reporting, interpreting physics information) (Lgr 11, pp. 129-130).*

Years 7-9 extends this core content with the same four general headings but with other, more advanced topics.

The Knowledge Requirements section of the curriculum policy document is important from the perspective of the assessment as it outlines the criteria for students to achieve grade levels from E through A in relation to the aims and content of the curriculum. It outlines, in general terms, the expectations for what students are able to do and demonstrate at the end of years 3, 6 and 9. In other words, this section sets out a blueprint for designing the assessment and we found that as we reviewed the actual national tests, we returned time and time again to this section of the curriculum for the purposes of comparison and validation.

The system of grades (from E up to A) that form the basis for this table are those used to report overall student achievement. However, as we shall describe later, its use for scoring individual assessment tasks and written test items can be seen as problematic.

Analyzing procedure

In this article we have made a text analysis focusing on tensions between what is stated in the tests in relation to the curriculum. Accordingly, in this research, we have analysed the Year 6 science curriculum, the corresponding national tests themselves, and the marking schemes associated with the tests and, based on this analysis, made some observations about the overall validity of the assessments. We analysed the biology, chemistry and physics curricula

and the tests (including the tests, the marking scheme and the information sent out to teachers) from the years 2013, 2014 and 2015 and found the design of them to be very similar to each other in all respects other than the specific scientific content.

For the purposes of this article, we have used the following model for analyzing the Year 6 national tests in science from 2013, 2014 and 2015. Assessment can be considered as a three-stage process involving (1) data gathering, (2) data evaluation, and (3) decision making. Since educational goals, as articulated in the curriculum, involve students' acquisition of knowledge, skills and attitudes, which cannot be observed directly, assessment is designed to gather evidence (data) from which inferences can be made about students' achievement of the curriculum goals. Planning any assessment therefore entails making decisions about:

1. What data to gather – student tasks or test questions represent the answers to this question;
2. How to evaluate the data – marking schemes or observation checklists represent answers to this question;
3. What to do with the result – the purposes of an assessment are revealed by the sorts of decisions made using the results.

In judging the validity of an assessment, one is considering the quality of the inferences being made about students' achievement of the educational goals and the appropriateness of the data selection and evaluation, in relation to the purposes for which the assessment is being conducted.

For this article, we illustrate our analysis with data from the physics curriculum and tests from 2013, with additional comments about the biology and chemistry as these are required. In addition, in our review of the national tests in physics, for this article, we illustrate our analysis of the entire tests through detailed accounts of one or two tasks per test section.

Findings

The year 6 national tests in physics are in three parts corresponding to the three aims of the curriculum and the three sets of knowledge requirements shown in Appendix 1.

Section A involves three tasks focused on physics and its social and environmental impacts. Consistent with the curriculum, the tasks are designed to enable pupils to demonstrate their

ability to ask questions, reason, use information and communicate in relation to energy, engineering, and the environment.

Section B involves six investigations in physics and is designed to enable pupils to demonstrate their skills in relation to research questions, experimental design, the execution of investigations, the collection and interpretation of data, and the critical appraisal of investigations, all using the areas of physics studied in these years (as noted above).

Finally, section C assesses pupils' understanding of physics concepts through a set of eighteen short-answer, multiple choice, and true/false questions covering the same range of physics topics.

Example 1: Pros and cons of wind turbines

One task in section A of the test concerned wind power and is typical of the tasks in this section. The pupils are given the following test item and are shown a short film about wind turbines – how they work and the advantages and disadvantages of their use.

> *It is becoming increasingly common to see wind turbines in the Swedish countryside. Both on TV and in magazines are discussions of wind power. There are many different opinions about wind turbines. Now you get to see a film about wind power in Sweden. You get to see the movie twice.*
>
> *Below is a conversation about wind turbines.*
>
> *'I think wind turbines are ugly.'*
>
> *'But that's just what you think. I think the main thing is that wind energy is renewable.'*
>
> *Your task is to prepare for the continuation of the conversation where you work out the arguments for and against wind power. Give as many arguments as possible. Try to deepen and broaden using your science skills.*
>
> *Keep in mind:*
>
> - *You should set out arguments both for and against wind power*
> - *You should write as many arguments as possible*
>
> *You should use your science skills*

This would appear to be exactly the sort of assessment task envisioned by the first aim of the curriculum and, in a live classroom; one can imagine a vigorous discussion following this film and task statement. There can be little question about the appropriateness of the data gathering phase of this assessment: the choice of responses from pupils to the given scenario is exactly the sort of assessment that is consistent with the first aim of the curriculum (as outlined above). In fact, one could imagine the discussion that might ensue if this task were undertaken in an open class situation. However, this assessment is not taking place in an open class situation but in an individual paper and pencil format which, while appropriate (some would say necessary) to obtaining individual achievement scores, nonetheless reduces the authenticity of the task as the 'conversation' is reduced to a monologue. This is the first point of tension – between curricular intentions and assessment choices – in this assessment; we shall return to it in our later discussion.

An even greater challenge to the validity of this section of the assessment arises from the data evaluation phase. The marking schemes for all three tasks have a number of significant common features. The system for evaluating students' work on each task is based on the same grading scheme (A – E & F) as is eventually used to describe overall student achievement. However, the scheme is in effect reduced to describing three of the passing grades (A, C and E). Finally, because making judgments about the quality of students' arguments (as implied by the words 'deepen' and 'broaden' in the curriculum and even in the task itself) also appears to be difficult and potentially unreliable. The passing levels (A, C, and E) are distinguished mainly by the number of arguments presented.

For example, the wind power task is evaluated as follows:

> *F: Pupil merely repeats what has been given in the task, gives a non-argument, or gives only one argument*
>
> *E: Pupil gives at least two arguments but only for or against*
>
> *C: Pupil gives at least three arguments, including both for and against AND at least one argument must be a science argument*
>
> *A: Pupil gives at least four arguments, including both for and against*

While the test paper reminds pupils to 'give as many arguments as possible' it does not warn them that this number will be almost the entire basis for evaluating their response. If pupil 1

elaborated (deeply and broadly) two arguments one in favour of wind power and one against, the best he or she could score would be an E grade, while pupil 2, who gives four short but different arguments, scores an A. Given the aim of the curriculum, one is forced to ask if this evaluation is a good reflection of what is intended by 'extending the dialogue and deepens or broadens it.' This is the second point of tension – between the 'knowledge requirements' specified in the curriculum and the marking scheme used in the assessment – to which we shall return.

Example 2: Investigating shadows

Section B of the assessment is focused on scientific investigations and the same two areas of tension can be seen in the assessment. First, since the assessment is mostly a written test, pupils' understanding of scientific inquiry and abilities to conduct investigations and experiments are being assessed through descriptions of and questions about aspects of investigations. This forces us to consider the degree to which the test questions measure students' achievement of the curriculum aims. Second, the marking schemes used by teachers to evaluate pupil responses are similar to those in section A and raise questions for us once again about how pupils' abilities can best be evaluated.

Section B is comprised of six parts, each touching on a different area of science content from the curriculum and each measuring pupils' abilities in relation to a different aspect of scientific inquiry, from the development of experimental questions, the critical review of experiments, decisions concerning appropriate data to collect in an experiment, designing an appropriate experimental methodology, and so on. For the purposes of this article, one of these tasks will be reviewed as typical. The task is as follows.

> *SHADOWS*
>
> *The pupils in a class have had an assignment to look at their own shadows. They stand in an open area outdoors on a clear sunny day in May. The shadow can be seen clearly all the time. They will look at their shadows every hour on the hour between 08.00 and 16.00.*
>
> *Help them to write three questions about the shadow properties and how it changes in different ways because the sun shines from different directions. They should get the answers to their questions by observing their shadows.*

Remember that questions beginning with 'why' and 'how' often cannot be answered by simply observing the shadow.

Write three questions following the example:

- *When is the shadow the shortest?*
- *...*

The marking scheme associated with this task gives examples of ineligible responses, examples of questions that need to be processed, and examples of questions that can be investigated directly. It also provides for just two grades, E and C, as follows:

- *For an E grade, pupils should give at least two questions dealing with the shadow properties - these may need to be processed further;*
- *For a C grade, pupils should give at least three questions, at least one of which could be investigated directly.*

We find, once again, a tension between curricular intentions and assessment choices. At first glance, the assignment could be a central part of a task that would fit the intentions stated in the curriculum that the students should be able to carry out systematic studies in physics. However, following the assessment scheme the students are only supposed to pose questions. Hence, they are expected to be writing down questions on a piece of paper, not carrying out any investigation. This is a second example of the first tension we described above.

Furthermore, as in the previous example, the quality of pupils' understanding is measured by the number of responses with, this time, the added distinction (only hinted at in the task) between questions that could be answered by direct observation and questions whose answers are based on observation but where further analysis is required.

Example 3: The C-part of the tests

The final section of the physics test comprises 18 questions about the concepts and theories of those areas of physics covered in this part of the curriculum. The test questions are of multiple-choice, sentence completion, or short answer type that can be readily evaluated by simple matching of the pupil's response to the (usually) limited number of correct responses. The same letter grades are applied once again to every question, although in this section, some

questions have only one correct response (and given an E grade), some have more than one (and given E and C grades), and some have multiple correct answers and three grades are used (E, C and A).

Aggregation of Scores

Before discussing the two substantive areas of tension that we have identified, we should note the method used to establish an overall grade for pupils' achievement.

- First, a pupil's score for each of the sections of the test is established by awarding points (5, 3, 1, 0) for each A, C, E or F grade achieved on each individual task or question, adding these up and expressing this total as a percentage of the maximum number attainable on that section.
- Second, these three percentage scores (for sections A, B & C) are averaged to provide an overall achievement score.
- Finally, this overall achievement score is converted back to a letter grade (A – E).

While this averaging system ensures an equality of importance of each of the three sections of the assessment, it also has the effect of magnifying the effect of any unreliability resulting from a vague, unclear or unevenly implemented marking scheme. For example we have noted in section A the relatively small difference in a pupil's performance on one task that could result in the achievement of an A or an E grade. This translates into a difference of 4 points out of a maximum of 15 points in section A, which converts to a difference of over 26% for section A and a difference of close to 9% in the overall achievement score which could easily convert to a full grade difference in the final grade. This raises a third point of tension in this assessment – between the intention of the evaluation system and its actual result – that we will discuss further.

Discussion

In general, the tests appear to be both innovative and very faithful to the curriculum. However, the task of evaluating students' responses to complex tasks and questions raises both practical challenges for teachers and interesting questions for researchers. Therefore, we want to conclude with some commentary on these and seek to situate them in the broader context of curriculum and assessment research in science education.

Our first impression on analysing the national tests is the high degree to which the tasks and items created for the tests correspond to the curriculum. This is not too difficult in relation to testing students understanding of science concepts and theories (as in section C). It is more difficult, as in section B, when attempting to assess students' abilities in relation to scientific inquiry (cf. Gott & Roberts, 2002; Harlen, 1999; Sund & Sund 2017). But it seems to be very difficult to create valid assessments in relation to the goals of science education known as Science, Technology, Society and Environment (STSE), represented here by the first curriculum aim and section A of the national test (cf. Orpwood, 2007).

However, three points of tension have been identified as we probed the test tasks and items further and examined the marking schemes attached to the tests. None of these should be taken as implying criticism of the test creators. Rather, we see them as unintended – and, we would argue, unfortunate – consequences of the way in which the overall assessment system was set up. These 'points of tension', as we have called them, resulted from the desire to create a national test of high validity, on the one hand, being constrained by aspects of the national curriculum/assessment/reporting system on the other.

The three points of tension identified in our accounts of the tests are:

1) Curricular intentions and assessment choices: The use of a paper and pencil test to assess student achievement of aims of education that call for pupil performances not well suited to such assessment.
2) The 'knowledge requirements' specified in the curriculum and the marking scheme used in the assessment: The use of a simple count of selected aspects of a performance, where the educational aim calls for a trained judgment of the quality of the performance.
3) The intention of the evaluation system and its actual result: The use of letter grades for evaluating pupils' performances on each task and item, the aggregation and conversion of these to numerical scores and the subsequent aggregation, averaging and conversion of the numerical scores back to letter grades.

Following the same way of thinking as in the analysing procedure, we will now discuss the issues of datagathering, evaluation and reporting and the consequenses for the validity of the tests.

Data gathering

The three aims of the curriculum call for very different types of learning on the part of students and these in turn call for equally different types of assessment. While nobody would assess the goal of swimming lessons with a written examination on the topic of swimming, it is all too common to find that the educational aim of pupils' developing the ability to conduct scientific investigations is assessed in this way. The alternative, performance assessment, where pupils are given real investigations to conduct, has been successfully used in national and international assessments, as well as at the school level and (cf. Sund & Sund, 2017), while there are challenges to overcome, this form of assessment is both more valid and authentic than any written tests about science investigation.

Similarly, aims of science education, in which pupils are expected to make, defend and refute arguments relating to the uses of science and technology in society and the environment, call for new forms of assessment. In particular, the first aim of physics education, as stated in the Swedish curriculum, appears to imply an interaction or conversation among pupils in which the quality of their participation should be the focus both of the instruction and the assessment. An individual, written test is therefore a pale shadow of what is implied in the curriculum. Assessment that involves an interactive performance is called for here. Law students use a moot court for exactly this form of assessment and there are many other contexts in which an individual participant's contribution to a group activity is the focus. Of course, such assessment is less common in the middle stages of education and therefore presents challenges. But it is very common in the early years, where teachers are trained to observe and evaluate children's work in a group setting. It is also very common in team sports, choirs, drama groups and other areas in which the individual is assessed but in a group setting.

The tensions arise here because, as is the case in most national testing systems, the decision to develop national tests tends to move into decisions about the sorts of items to place on a written test and the intermediate stage of questioning the mode(s) of assessment most suited to the goals can be passed over. Yet we would argue that this is one of the most important questions to be answered if the goal of a valid assessment is to be achieved. If, in addition, one of the primary purposes of putting the national tests in the database to model good assessment practice to teachers, then this choice of assessment mode is of even more critical importance.

Data Evaluation

However well the tasks in a test are aligned with the aims of the curriculum, the ways in which pupils' performances on those tasks are evaluated can affect the validity and reliability of an assessment. We have highlighted examples from the physics tests where pupils' abilities at 'reasoning' were evaluated simply by the number of acceptable reasons given rather than the quality of their arguments.

A full discussion of appropriate modes of evaluating such a complex ability as reasoning is beyond the scope of this article. But an important distinction can be made between quantitative evaluations, where the evaluator counts correct responses, and qualitative evaluations, where the evaluator makes judgements about performances. Such a distinction is obvious in the Winter Olympics: while ice hockey and cross-country skiing are evaluated quantitatively (based on goals or times), figure skating and ski jumping are evaluated qualitatively (by trained judges). Ice hockey is not evaluated by the quality of the players' skating or figure skating simply on the number of jumps. Rather the modes of evaluation are carefully matched to the essential nature of the participants' performances.

The same careful matching should be an essential part of developing a valid and reliable assessment in education also. Once again, this is not always easy in practice. International assessments use trained teams of markers to make those judgments based on a carefully tested rubric. If evaluations are to be carried out at the local level, as opposed to centrally, then teachers require training and sets of exemplars to develop their ability to make such judgments of performance quality. The use of an apparently more reliable but inappropriate quantitative evaluation system, however, can undermine the validity of an otherwise promising task. In addition, the form of evaluation to be used needs to be signalled clearly to the pupils so that they understand exactly what is expected of them.

Data reporting and use

The final point of tension that was apparent in our analysis of the national tests in science has less to do with the substance of the curriculum or assessment but everything to do with how the qualitative and quantitative evaluations of pupils' performances on individual tasks and test items are combined into overall reports of achievement in science.

Several questions need to be discussed if the tests are used in everyday practice: What is the most appropriate way to score each task and test item? How should these scores be further combined? How should a pupil's overall abilities in physics best be represented? What uses will these overall representations serve and for whom? We do not have answers to these questions but we believe that they are all in need of consideration.

Concluding remarks

In this study, we have clarified and discussed three different tensions between the curricula and the Swedish national tests in Biology, Chemistry and Physics. Our impression is that the content of the tests is clear, relevant and corresponds well to the goals and knowledge requirements stated in the curriculum. However, the tensions identified and discussed there have consequences for the validity of the tests. For example, if we want to assess the students' ability to participate in a discussion, it is not enough to ask them just individually to write down as many arguments as possible in a paper and pencil test. Hence, if these tests are used in teachers' assessment practice, they need to be complemented with other forms of testing as well. We think that the tests, and especially the design and content of the questions, can be a good inspiration for teachers and help them assess parts of their students' knowledge. But, given the tensions described here and the constraints following the format of paper and pencil tests, it is necessary to regard these items as only one of many resources teachers can use in their assessment practices. Based on this case study, we point to the fact that in order to assess the students' knowledge with high validity, teachers need to gather other kinds of data than data from paper and pencil tests. In further development and use of tests in science education, and in other subjects as well, the test constructers should consider these tensions. We want to stress that, since not all kinds of competencies can be tested in these kinds of tests, it is important to further discuss the tests in the Assessment Portal and what they can be used for.

Acknowledgment

The authors would like to thank professor Jim Ryder and the members of the Research Group for Comparative Didactics at Uppsala University for valuable comments on the text.

This work was supported by The Swedish Research Council (Vetenskapsrådet) under Grant 2012-5769

References

Almqvist, J. & Lundqvist, E. (2013). De nationella provens innehåll: Vilken scientific literacy mäts i NO-proven. In E. Lundqvist, R. Säljö och L. Östman (Eds.) *Scientific literacy*. Malmö: Gleerups Utbildning AB.

Au, W. (2007). High-stake testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), *258-267*. https://doi.org/10.3102/0013189X07306523

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215-232. https://doi.org/10.1080/09695941003696016

Broadfoot, P. & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy & Practice, 11*(1), 7-27. https://doi.org/10.1080/0969594042000208976

Bybee, R., McCrae, B. & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of research in science teaching, 46*(8), 865-883. https://doi.org/10.1002/tea.20333

Cimbricz, S. (2002): State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2). http://dx.doi.org/10.14507/epaa.v10n2.2002

Englund, T. (1986). *Curriculum as a political problem. Changing educational conceptions, with special reference to citizenship education*. Lund: Studentlitteratur/Chartwell-Bratt.

Fensham, P. (1988). Familiar but different: some dilemmas and new directions in science education. In P. Fensham (Eds.), *Development and dilemmas in science education*, (pp. 1-16). London: The Falmer Press.

Gott, R. & Duggan, S. (2002). Problems with the assessment of performance in practical science: Which way now? *Cambridge Journal of Education, 32*(2), 183-201. https://doi.org/10.1080/03057640220147540

Harlen, W. (1999). Purposes and procedures for assessing science process skills. *Assessment in Education: Principles, Policy & Practice, 6*(1),129-144. https://doi.org/10.1080/09695949993044

Hirsh, Å. (2016) Nationella prov i grundskolan. En studie av hur lärare och rektorer uppfattar och hanterar prov och provresultat. [National tests in compulsory school. A study of how teachers and school leaders perceive and handle tests and test results.]. Stockholm: Skolverket.

Lgr 11. Curriculum for the Compulsory School, Preschool class and the Recreation Centre, 2011.Stockholm: The Swedish National Agency for Education.

Lidar, M., Lundqvist, E., Ryder, J. & Östman, L. (in review). The transformation of teaching habits in relation to the introduction of grading and national testing in Sweden. Submitted to Research in Science Education.

Lidar, M., Karlberg, M., Almqvist, J., Östman, L. & Lundqvist, E. (2017). Teaching Traditions in Science Teachers' Practices and the Introduction of National Testing. *Scandinavian Journal of Educational Research*. Published online: 26 Apr 2017. https://doi.org/10.1080/00313831.2017.1306802

Linder, C., Östman, L., Roberts, D. A., Wickman, P.-O., Erickson, G. & MacKinnon, A (Eds.)(2010). *Exploring the landscape of scientific literacy*. London: Routledge

Lundqvist, E., Almqvist, J., & Östman, L. (2012). Institutional traditions in teachers‟ manners of teaching. *Cultural Studies of Science Education, 7*(1), 111-127. https://doi.org/10.1007/s11422-011-9375-x

Lundqvist, E. & Sund, P. (2016). Selective traditions in group discussions: Teachers' views about good science and the possible obstacles when encountering a new topic. *Cultural Studies of Science Education.* Published online 25 November 2016. https://doi.org/10.1007/s11422-016-9768-y

Marty, L; Almqvist, J. & Venturini, P. (2017). Teaching traditions in science education in Switzerland, Sweden and France: A comparative analysis of three curricula. *European Educational Research Journal.* Published online 14 June 2017. https://doi.org/10.1177/1474904117698710

Messnick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. https://doi.org/10.1002/j.2333-8504.1992.tb01470.x

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149-170. https://doi.org/10.1080/09695940701478321

Newton, P. E. & Baird, J.-A. (2016). The great validity debate. *Assessment in Education: Princiles, Policy and Practice, 23*(2), 173-177. https://doi.org/10.1080/0969594X.2016.1172871

Orpwood, G (2001). The role of assessment in science curriculum reform. *Assessment in Education: Principles, Policy & Practice, 8*(2), 135-151. https://doi.org/10.1080/09695940125120

Orpwood, G. (2007). Assessing Scientific Literacy: Threats and Opportunities. In C. Linder, L. Östman, & P.-O.Wickman (eds.) *Promoting Scientific Literacy: Science Education Research in Transaction, Proceedings of the Linnaeus Tercentenary Symposium, Uppsala University, Sweden, May 2007*. Uppsala: Uppsala University, 2007, pp. 120-129.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265-279. https://doi.org/10.1016/j.tsc.2013.07.006

Östman, L. (1996). Discourse, discursive meanings and socialization in chemistry education. *Journal of Curriculum Studies, 28*(1); 37-55. https://doi.org/10.1080/0022027980280102

Roberts, D. A. (1982). Developing the concept of "curriculum emphases" in science education. *Science Education, 62*(2), 243-260. https://doi.org/10.1002/sce.3730660209

Roberts, D. A (2007). Scientific literacy/science literacy. I S. K. Abell & N. G. Lederman (Eds.). *Handbook of research on science education* (pp. 729-780). Mahwah, NJ: Lawrence Erlbaum.

Ryder, J., Lidar, M., Lundqvist, E. & Östman, L. (in review). Expressions of agency within complex policy structures: Science teachers' experiences of educational policy reform in Sweden. Submitted to International Journal of Science Education.

Sampson, V. & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education. Current perspectives and recommendations for future directions. *Science Education, 92*(3), 447-472. https://doi.org/10.1002/sce.20276

Sund, P. & Sund L. (2017). "Alla gör fel?!" – Hinder för lärares bedömning av elevers praktiska förmågor under ett nationellt prov. *NorDiNa, 13*(1), 3-16. http://dx.doi.org/10.5617/nordina.2845

Skolverket (2015a). Lärarinformation, Ämnesprov, läsår 2014/2015, Fysik årskurs 6. Stockholm: Skolverket.

Skolverket (2015b). Skolreformer i praktiken. Hur reformerna landade i grundskolans vardag 2011-2014. [School reforms in practice. How the reforms landed in the primary school practice 2011-2014] Rapport 418:2015. Stockholm: Skolverket.