The development of a national e-test on science competence for the third school level: An assessment to support learning

Katrin Vaino[a1], Triin Rosin[a], Ülle Liiber[a, b], Regina Soobard[a], Moonika Teppo[a], Ana Valdmann[a], Elle Reisenbuk[c], & Miia Rannikmäe[a]

[a] Centre for Science Education, Institute of Ecology and Earth Sciences, University of Tartu
[b] Geography Department, Institute of Ecology and Earth Sciences, University of Tartu
[c] Estonian Education and Youth Board

Summary

Theoretical framework

The contemporary concept of science competence includes scientific knowledge and its application, inquiry skills, creativity, the ability to solve problems and make well-considered decisions (OECD, 2020), sustainable values, collaboration and communication skills, and the ability to self-regulate one's learning (Choi et al., 2011).

In Estonia, since 2006, the PISA survey has been conducted among 15–16-year-old students, including an assessment of their science competence. However, its results do not provide feedback on the level of competence of an individual or group of students, nor specifically on the level of science competence defined in the Estonian national curriculum for basic schools (2023).

The given circumstances necessitated the need for a nationwide test based on the Estonian national curriculum, which would, on the one hand, provide objective and comparable feedback about the level of Estonian students' scientific competence to assess the effectiveness of the educational system and, on the other hand, provide specific feedback to the student and the teacher (Ministry of Education and Research, 2014).

Education internationally has moved towards competence-based curricula, which in turn requires a changed approach to assessment, which focuses on supporting teaching and learning and developing a self-directed learner (Assessment Reform Group, 1999). This trend has also increased interest in

[1] Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, 51003, Tartu; katrin.vaino@ut.ee.

the potential and use of diagnostic (low stake) tests. The new generation of diagnostic tests can also offer constructive feedback that defines the student's existing knowledge, skills, attitudes, and areas for improvement (Csapó & Molnár, 2019). Therefore, this type of test is often based on the sociocultural theory (*Ibid.*), according to which learning is possible if it occurs in the student's zone of proximal development (Vygotsky & Cole, 1978; Daniels, 2007).

Therefore, this study aims to develop a nationwide e-test that is able to assess the level of science competence achieved by students at the end of the third school level and provide feedback to the student, teacher, parent, and, in a generalised manner, to the educational community. This paper aims to describe the design process, explain the various design decisions made throughout (e.g. how the users – students interact with the created design), and, in more detail, the resulting final design (e-test).

The conceptual framework of the science e-test is based on the report of Pedaste et al. (2017), including the contemporary definitions of science competence as introduced above. The report presents a potential assessment model and the initial procedure for preparing and executing the task. Accordingly, the test should assess students' science knowledge and inquiry, communication, and problem-solving skills via ten characteristics aligning with the science curriculum as part of the Estonian national curriculum for basic schools (2023). According to the report (Pedaste et al., 2017), the test should consist of context-based tasks, the uses of which should help, amongst other benefits, to maintain the student's interest, as the absence of their interest poses a serious threat to the validity of the conclusions drawn from the test results (Finn, 2015).

Methodology

The test was developed in cooperation with the University of Tartu and the Education and Youth Board (Republic of Estonia). In this study, the design-based research approach was used, which tries to bridge the gap between educational research and actual school practice to increase the viability of the developed design (van der Akker et al., 2006) and to improve the existing educational practice through systematic, flexible and repeated analysis, further development, implementation, and correction (Cobb et al., 2003). As a result of the study, diverse data were collected to assess the appropriateness of the test design. Due to the complexity of the process, only a part of the results is presented in this paper. The characteristics of the final design are described in more detail.

The test development process in each of the five design cycles consisted of four steps (adapted from Reeves, 2006):

(1) Problem analysis by both researchers and practitioners, goal setting of activities.

(2) Development of the test tasks and the format of feedback.

(3) Test evaluation and piloting (see Table 1).

Table 1. The number of test tasks and the number of participating students according to the language of performance and the type of school

Year	Grade	Number of tasks	Number of participating students according to the language of performance and the type of school
2018	9.	51	1226 (Estonian); 274 (Russian); all general education
2019	9.	37	539 (Estonian); 269 (Russian); all general education
2020	9.	35	108 (Estonian); general education
2021	10.	35	1363 (Estonian); 453 (Russian); general education 1323; vocational education 493
2022	10.	35	532 (Estonian), 260 (Russian), general education 713, vocational education 79

(4) Analysing, reflecting, and documenting the results. The results of the prototype test were analysed using (i) descriptive statistics from 2019, (ii) exploratory and confirmatory factor analysis, and (iii) Rasch analysis based on item response theory (IRT) illustrated with Wright maps. Correspondence of individual items to specific levels (base, middle, high, or top) was derived from Wright's maps. Based on the nature of the items at each level and the respective scores of the items, the learning outcomes per level were described.

Results and discussion

As in the case of a low-stakes test, the validity of conclusions may be jeopardised by students' low interest in solving the tasks (Finn, 2015). This aspect was particularly examined in the study. It has been found that, over the years, two-thirds of respondents have perceived the test as interesting. One potential way to increase the response of uninterested students would be to enrich the tasks with media-rich stimuli (Boyle & Hutchison, 2009) in future tests.

Additionally, over the years, students have perceived the test as rather challenging, which can be explained by the fact that solving tasks requires transferring acquired knowledge into everyday life contexts. Another explanation could be that the assessment of inquiry skills has not yet become a self-evident part of science learning, as indicated earlier by Henno et al. (2017).

On the other hand, the number of students perceiving the test as difficult has decreased over time. This may be attributed to both the fact that the test itself has started to impact teaching and assessment methods in schools (Rosin et al., forthcoming) and because the number of tasks in the test has decreased, reducing the potential cognitive overload for students.

While reducing the test difficulty was not a direct goal for the team to ensure the test's discriminative power, monitoring the perceived difficulty of the tests by students in future administrations is advisable.

Based on exploratory factor analysis of the test results, a multifactorial solution was hoped to be found to reduce the number of variables to explain and interpret the results. In reality, based on exploratory and confirmatory factor analysis of several test results (2019–2022), multifactorial models were obtained, which had relatively good fit indicators but whose factors were essentially uninterpretable. Therefore, it was decided to stick to a one-factor solution. According to Cronbach's alpha coefficient, the internal consistency of the factor items was 0.89 (final test), which could be considered "good." Therefore, this study proposes a one-dimensional solution; the latent trait behind the factor items can be called "science competence."

The final assessment model, which successive versions have guided the development of the whole test, includes nine attributes, which are categorised into four components: subject knowledge, inquiry skills, problem-solving skills, and communication skills (see Harno, 2023b). The model includes a description of the learning outcomes characteristic of both – the updated Bloom's (Anderson & Krathwohl, 2011) and SOLO taxonomy (Biggs & Collis, 2014). The latter was primarily used to describe learning outcomes collected through open-ended responses.

The final version of the test consists of four sub-tests, each representing one context ("story") and a total of 35 items designed to determine the level of students' science competence, described by nine attributes, and for the sake of clarity, grouped into four categories (see Table 2).

Table 2. Assessed components, respective attributes, the number, type and examples of the items

Components and attributes	Open-ended	Short answer	Multiple choice	Total	Examples
1. Science knowledge					**The student:**
Explaining natural phenomena			3	3	Arranges natural processes in the correct order.
Use of scientific concepts, symbols, and units		1	4	5	Selects appropriate units for the given situation.
Understanding of scientific models	2		4	6	Drags concepts to the concept map. Explains the model in their own words.
2. Inquiry skills					
Posing a research question or hypothesis			2	2	Selects a research question based on the given situation; formulates a hypothesis based on the graph.
Designing experiment and evaluating its quality	3		2	5	Arranges the stages of the experiment based on the given situation; explains in their own words why the described experiment falls short in quality.
Analyzing information and drawing conclusions	1	1	2	4	Analyzes data presented in the form of a table or graph and draws conclusions.
3. Problem-solving and decision-making					
Problem-solving and decision-making	3		4	7	Explains the nature of the problem based on the given situation and organizes activities to eliminate the malfunction in the described system; considers a dilemma between two social decision alternatives and justifies the decision, including expressing personal values.
4. Communication skills					
Finding information and assessing its reliability	1		1	2	Searches for relevant information on the internet; evaluates the reliability of the given source, highlighting aspects that support their evaluation.
Understanding of and writing science text	1			1	Summarizes textual and visual information in their own words (50–60 words).

The provided feedback informs the student of their level of science competence (zero, base, medium, high, top) and, according to the level, describes their existing competence through nine assessed attributes. In addition, the feedback provides recommendations for setting the next learning goals – these are posed one step higher than the achieved level. The teacher receives feedback on the level of both individual and groups of students that enables him/her to identify students' learning needs more effectively and, based on that, modify his/her teaching.

Keywords: science competence, diagnostic e-test, feedback, design-based research, context-based