International Journal of
STEM Education

RESEARCH Open Access

Towards a science inquiry test in primary education: development of items and scales

Margus Pedaste[1*], Aleksandar Baucal[1,2] and Elle Reisenbuk[3]

Abstract

Background: Inquiry-based learning is widely applied in science education; however, so far, the outcomes of learning process have been systematically assessed mainly at the secondary school level. For primary school students, there is no valid instrument for assessing the outcomes of their science inquiry. The aim of the current study was to develop a test for assessing science learning outcomes (analytical skills, planning skills, interpretation skills, and science knowledge) related to the five phases of inquiry-based learning (Orientation, Conceptualization, Investigation, Conclusion, and Discussion) at primary education level (ISCED 1).

Results: A set of contextualized science tasks was created to assess each of the learning outcomes at three levels. The Science Inquiry Test for Primary Education (SIT-PE test) was developed through several phases, including pilot studies with large groups of fourth-grade students (10 to 11 years of age). The 1 PL Item Response Theory model was used to analyze the quality of the test and items based on the test's reliability score, item difficulty measure, infit and outfit indices, estimation of item discrimination, item-scale correlation, and the quality of the scoring key. The final test, consisting of 24 items, was used with a sample of 1868 students. The analysis showed the SIT-PE test to be of good quality on test level and item level and to also have good predictive validity. Confirmatory factor analysis revealed that the correlated factors model and second-order factor model of the science learning outcomes both had a good fit to data collected with the SIT-PE test. Confirmatory factor analysis confirmed the multidimensionality of science learning outcomes and validated four dimensions of the model: analytical skills, planning skills, interpretation skills, and science knowledge.

Conclusions: In conclusion, the SIT-PE test could be further used for assessing students' inquiry competence in primary education. However, it could be even further improved in several ways and this study provides guidelines on how to do that. In addition, the SIT-PE provides test developers with information on how to design derivations of the SIT-PE test for assessing particular science inquiry outcomes or the same outcomes in older age groups as well.

Keywords: Inquiry-based learning, Performance assessment, Primary education, Science, Assessment, Item response models, Confirmatory factor analysis

* Correspondence: margus.pedaste@ut.ee
[1]University of Tartu, Salme 1a, 50103 Tartu, Estonia
Full list of author information is available at the end of the article

Introduction

Inquiry-based learning is one of the main approaches to learning science (see, e.g., National Research Council, 2000; Osborne & Dillon, 2008). It has been effective—in comparison with more teacher-centered methods—for the conceptual understanding of science, but also for enhancing students' motivation and interest in learning science (see, e.g., Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Constantinou, Tsivitanidou, & Rybska, 2018; Furtak, Seidel, Iverson, & Briggs, 2012). Areepattamannil, Cairns, and Dickson (2020) have found, through the use of OECD PISA data (Organization for Economic Cooperation and Development, Program for International Student Assessment) of 428,197 students from 66 countries, that inquiry-based teaching significantly positively predicts students' enjoyment of science, interest in broad science topics, motivation to learn science, and science self-efficacy. Therefore, inquiry-based learning has been widely applied in the context of science education, but it is applicable in many other disciplines as well. For example, on a general level, Keselman (2003) describes inquiry-based learning as an educational strategy where learners construct knowledge using the methods applied by scientists. Many authors have emphasized that inquiry-based learning is a process where learners are actively involved (see Constantinou et al., 2018; de Jong & van Joolingen, 1998; Mäkitalo-Siegl, Kohnle, & Fischer, 2011; Pedaste, Mäeots, Leijen, & Sarapuu, 2012). This means that in doing inquiry, students also need to plan, monitor, and evaluate their learning process (De Jong, Kollöffel, van der Meijden, Kleine Staarman, & Janssen, 2005). This shows that the inquiry approach used in learning science is also applicable for improving more generic learning skills: how to plan activities, how to monitor their progress, and how to evaluate outcomes and make changes in the initial plan if needed.

In spite of the wide use of the inquiry-based learning process as a learning approach (see Archer-Kuhn, Lee, Hewson, & Burns, 2020; Liu, Zowghi, Kearney, & Bano, 2021; Misra, 2020), few studies have focused on the assessment of the outcomes of the inquiry process. Both inquiry skills and knowledge acquired through inquiry could be regarded as outcomes of the inquiry process. Some earlier studies might have focused on one or another skill necessary in the inquiry process, but few studies have covered all dimensions of the inquiry process. Therefore, it is first important to specify more clearly what the inquiry skills are that could be differentiated empirically. Second, we need to find how to assess both the inquiry skills and science knowledge necessary or acquired in the inquiry process. It seems that these questions are especially understudied in the context of primary education, where the inquiry approach has often been used without specific assessment instruments. Therefore, we can see that there is a need for developing a test that could be used for assessing the outcomes of the inquiry process at the primary education level. The next chapters of the article open up the concept of inquiry-based learning and how the outcomes of inquiry-based learning have been assessed. This literature review is the basis for creating the conceptual framework of our study as introduced in the chapter following the literature review.

Literature review

Inquiry-based learning is a complex process; it is usually guided by introducing different phases that structure inquiry. Pedaste et al. (2015) conducted a systematic literature review and found descriptions of inquiry phases in 32 articles, which used 109 different terms. By removing the overlapping terms, sequencing the phases, and organizing these into larger groups, five general inquiry phases were identified: Orientation, Conceptualization, Investigation, Conclusion, and Discussion. Orientation is a process of getting to know about a situation and addressing a learning challenge through a problem statement. Conceptualization is for defining the problem stated in the Orientation phase and for conceptualizing it by formulating research questions and/or hypotheses. In the Investigation phase, learners need to design a plan for finding an answer to the research question or to obtain evidence to accept or reject the hypothesis. In its planning sub-phase, learners need to specify all the steps necessary for collecting data, the equipment and materials for collecting data, and how to ensure the validity and reliability of the data collection. Next, they need to collect data according to the plan. After, they must analyze and interpret the data. In the Conclusion phase, the outcomes of the Conceptualization and Investigation phases will be combined to draw conclusions. The Discussion phase is conducted either at the end of the whole inquiry process or in parallel with all other phases—the learners could discuss the process and outcomes of the Orientation, Conceptualization, Investigation, and Conclusion phases. It would be best to do this in every phase before proceeding to the next because it helps the acquisition of inquiry skills by reflecting on the process and getting feedback from others. Thus, it is important to mention that according to the synthesis of all frameworks found in the systematic literature review, the inquiry process is not linear: it can start from different phases, and some of these phases, e.g., the Discussion phase, could be repeated several times in an inquiry process. In addition, it was found that whereas the Orientation, Conceptualization, Investigation, and Conclusion phases focus mainly on the problem to be solved by the inquiry approach, the Discussion phase goes beyond that, leading the learners more towards the analysis of the inquiry process, not only to its outcomes. Therefore, the Discussion phase is especially valuable in acquiring inquiry skills through reflection on the process (see Liu et al., 2021 for a review of studies focusing on collaborative inquiry).

The inquiry framework developed by Pedaste, Mäeots, et al. (2015) is widely applied in learning sciences. For example, only recently, several guided inquiry-based learning environments such as the Ark of Inquiry (see Pedaste et al., 2015) and the Go-Lab (see De Jong, Sotiriou, & Gillet, 2014) have been introduced. These environments apply the inquiry cycle of Pedaste, Mäeots, et al. (2015) and have been used by more than 300,000 students and teachers across the globe. However, these studies do not provide instruments for testing inquiry skills. This framework integrates wide knowledge about inquiry frameworks described in 32 research papers regarding definitions and cycles of inquiry. Therefore, this synthesis could be taken as a widely used framework about inquiry-based learning that integrates other well-known views on inquiry.

The aforementioned inquiry-based learning framework continues to be widely used to guide the design of science learning and science teacher education (see Kuter & Özer, 2020; Oguz & Aybars, 2019; Wu & Wu, 2020). However, only a few studies have focused on the assessment of students' science skills according to either this framework or any other inquiry-based learning framework that has been used to synthesize the framework of Pedaste, Mäeots, et al. (2015). Even less attention has been paid to assessing the inquiry process outcomes on a primary school level (ISCED 1). For example, Schiefer, Golle, Tibus, and Oschatz (2019) developed an instrument for assessing 8- to 10-year-old students' understanding of the phases of the inquiry cycle. They confirmed the reliability of their 15-item test using Item Response Theory (IRT). However, their test focused on the assessment of the students' understanding of the typical order and necessity of the steps of the inquiry cycle and not on the skills needed in every phase or the knowledge necessary in the inquiry process or acquired as an outcome of inquiry. Some other studies have focused on assessing students' skills, but only in one or another inquiry phase. For example, Pöntinen, Kärkkäinen, Pihlainen, and Räty-Záborszky (2019) focused on a small group of 11- to 12-year-old students to understand their questioning skills. Arini, Suratno, and Yushardi (2019) assessed students' communication skills in inquiry learning. Shanks et al. (2017) focused on measuring experimental design ability. Several issues of the online testing of science inquiry have also been identified by DeBoer et al. (2014), e.g., online testing's usability, effectiveness, and comparability with paper-and-pencil testing. More specifically, DeBoer et al. (2014) focused on comparing the effect of different modalities on measuring science knowledge and skills; they found that interactive online modality that allows students to make connections between the objects enables testing more complex reasoning skills than simply showing animations or images the students cannot interact with. Therefore, the online testing of inquiry skills is definitely worth further studies.

One of the most advanced and widely used instruments following the inquiry-based learning framework has been developed in the USA for measuring scientific inquiry based on a data set from the Virtual Performance Assessment (Scalise & Clarke-Midura, 2018). A Virtual Performance Assessment task engages students in a guided inquiry through problem-solving, modeling, and exploration. The case study by Scalise and Clarke-Midura (2018), based on an IRT analysis, showed that this online instrument could be used to describe student proficiency in scientific inquiry with respect to two general science skills: inquiry and explanation. Inquiry included posing questions, designing investigations, and carrying out the investigation. Explanation included analyzing the results of inquiry, drawing conclusions, and communicating results. Thus, they omitted the Orientation phase and only partly covered the Conceptualization and Discussion phases. In addition, their study had an important limitation: they used only one inquiry task in a very specific context. Finally, they approached older students (ISCED 2).

Similarly, Kuo, Wu, Jen, and Hsu (2015) have developed and validated a multimedia-based assessment instrument of the inquiry abilities of secondary school students. Their test consisted of 101 items in 29 tasks, which might be too many for primary school students. The inquiry abilities differentiated in their study were, indeed, quite similar to the ones identified in our test. They specified four abilities: questioning, experimenting, analyzing, and explaining. However, they found that several items did not have an acceptable fit or had an unacceptable discrimination index. One more issue found by them was that although it was possible to distinguish the four inquiry abilities, the correlation coefficients between them ranged from .87 to .96. Thus, the different inquiry abilities are strongly related to each other.

The same age-specific limitation has occurred in implementing large-scale international assessments, and the dimensionality of inquiry skills has not been reported about these tests according to our best knowledge. For example, inquiry skills have also been highlighted in the international OECD PISA test, which focuses on middle school students. In contrast, the IEA TIMSS test (International Association for the Evaluation of Educational Achievement Trends in International Mathematics and Science Study) (Jones, Wheeler, & Centurino, 2015; OECD, 2013) focuses on fourth-grade students' inquiry skills. The TIMMS2019 Science Framework distinguishes between content domains and cognitive domains (Mullis & Martin, 2017). The cognitive domains are knowing, applying, and reasoning; in reasoning, students need to analyze, synthesize, formulate questions, hypothesize and predict, design investigations, evaluate, draw conclusions, generalize, and justify. This is very much in line with the inquiry framework of Pedaste, Mäeots, et al. (2015), although only 20% of the TIMSS test is designed to assess reasoning; however, differentiation of these dimensions as factors has not been tested with factor analysis. In addition, there is one more limitation in the PISA and TIMSS tests—they do not clearly follow the inquiry phases, although they claim to focus on inquiry activities. Thus, these tests show that inquiry is important in the context of science skills worldwide, but they do not focus specifically on an inquiry framework or science knowledge in the context of inquiry tasks.

Conceptual framework and research questions of the study

The analysis of inquiry-based learning frameworks and different tests developed for assessing the outcomes of the inquiry process led us to the formulation of the conceptual framework for the current study. We found that the inquiry framework of Pedaste, Mäeots, et al. (2015) synthesizes many frameworks describing the inquiry process and, therefore, represents a wide view on inquiry-based learning. Therefore, we decided to proceed from this study in developing our conceptual framework. However, we also found that it would be important to add science knowledge as a dimension of the outcomes of an inquiry process. Accordingly, we proposed that the outcomes of scientific inquiry could be conceptualized through inquiry skills and scientific knowledge. Our aim was to develop a test suitable for assessing different inquiry skills and scientific knowledge. However, we were not sure what kind of skills could be differentiated in the case of primary school students, because even though inquiry is considered important in science learning, it has not been systematically assessed on primary education level using an instrument that could be widely applied in different contexts. Thus, the current study aimed at filling this gap by developing a Science Inquiry Test for Primary Education (SIT-PE) based on the inquiry-based learning framework developed by Pedaste, Mäeots, et al. (2015). The focus of the paper is on developing the SIT-PE test and assessing its quality in order to provide answers to the following research questions:

1. What is the quality of each individual item of the SIT-PE test and the potential of the test to measure science learning outcomes?
2. What are the latent variables that can be differentiated with the SIT-PE test?
3. What is the predictive validity of the SIT-PE test?

Methods

Developing the test items and compiling the test

The development of the SIT-PE test consisted of four phases. In the first phase (see Fig. 1), an expert group consisting of science education researchers, science teacher educators, and science teachers created a list of inquiry-based learning outcomes that followed the framework developed by Pedaste, Mäeots, et al. (2015): analytical skills (mainly needed in the Orientation, Conceptualization and Investigation phases), planning skills (mainly needed in the Investigation phase), interpretation skills (mainly needed in the Conclusion and Discussion phases), and necessary knowledge on the relevant topics of inquiry. More information about the connection between the learning outcomes and inquiry framework is presented in Table 1. The connection of the phases to science knowledge is not introduced in the table, because the knowledge assessed by the test was not specifically linked to any of the inquiry phases. The knowledge assessed with SIT-PE was topic-specific content knowledge.

In the second phase, the initial version of the SIT-PE test was developed by the same expert group that worked in the first phase. For that, the science teachers developed a number of problem-solving tasks following the inquiry cycle and the authors of the article selected the tasks that were well in line with the conceptual framework of the current study. Next, the items of the selected tasks were revised so that they would allow differentiating analytical skills, planning skills, interpretation skills, and science knowledge. Finally, the SIT-PE test was used in a national assessment as the first pilot study (N = 589). In this study, the difficulty of the test items (average percentage of the maximum score the students gained for each item), evaluation of students' test-taking motivation, assessment of the test's difficulty, and teachers' feedback were used to develop recommendations for improving the test. Next, the test items were revised and used in the first main study (N = 1,798), where, again, the difficulty of the items was analyzed and students' and teachers' feedback was collected.

In the third phase, the findings from the second phase were used to revise the test items and create additional ones to compile two versions of the test: one with multiple-choice questions for all items where it was possible and one with open-ended questions to test the difficulty levels of different types of items and their quality. The second pilot study (N = 287) aimed at decreasing the number of tasks and items and reducing the time needed to complete the test (by identifying the questions

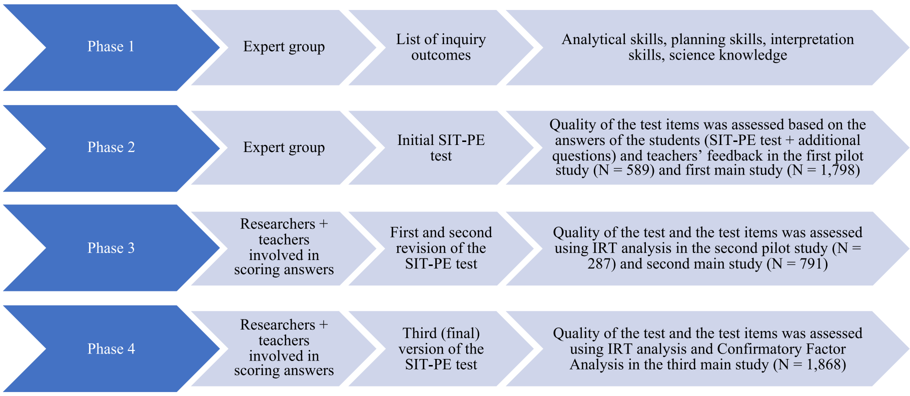

Phase 1
Expert group
List of inquiry outcomes
Analytical skills, planning skills, interpretation skills, science knowledge
Phase 2
Expert group
Initial SIT-PE test
Quality of the test items was assessed based on the answers of the students (SIT-PE test + additional questions) and teachers' feedback in the first pilot study (N = 589) and first main study (N = 1,798)
Phase 3
Researchers + teachers involved in scoring answers
First and second revision of the SIT-PE test
Quality of the test and the test items was assessed using IRT analysis in the second pilot study (N = 287) and second main study (N = 791)
Phase 4
Researchers + teachers involved in scoring answers
Third (final) version of the SIT-PE test
Quality of the test and the test items was assessed using IRT analysis and Confirmatory Factor Analysis in the third main study (N = 1,868)

Fig. 1 Four phases of designing the SIT-PE test

that would be most suitable for differentiating the three levels in inquiry-based learning outcomes and by selecting the multiple-choice questions, if possible) to be used in the second main study (N = 791). The second main study was conducted for assessing the quality of the final test items by describing the difficulty and quality measures of the items. The difficulty of the test items was described based on the item difficulty measure found in IRT analysis. The distinction between the levels was made on the basis of the WrightMap, which showed where the more distinct differences were between item difficulty measures. The qualitative differences between the items on different levels were described by analyzing the content of the items on a particular level. The descriptions were used in providing students with feedback on their inquiry outcomes and recommendations for the further learning process. The fourth phase assessed the quality of the final test. This is the focus of the current article and will be discussed in the following sections.

Sample of the study

In this article, we use data collected in the fall of 2018 in phase 4 of developing the SIT-PE test. The sample of the study consisted of schools that voluntarily agreed to the assessment of their students' inquiry learning outcomes. The students were not motivated by someone else to participate in the test. They did not get a grade based on the results of the tests, and no comparisons of the results were made on the student or school level. No incentives were offered to either the schools or students participating in the testing. In 2018, the final test was completed by 1868 students from the fourth grade (9 to

Table 1 Association of inquiry-based learning outcomes and inquiry skills

Inquiry phase	Analytical skills	Planning skills	Interpretation skills
Orientation	Analyzing the problem situation to extract key variables and to state the problem	Planning the procedure of getting acquainted with the problem situation	Interpreting the problem situation in a personal context to make it more meaningful
Conceptualization	Analyzing the problem statement to specify relevant knowledge to formulate research questions or hypotheses consisting of dependent and independent variables	Planning the procedure of analyzing the problem statement and collecting relevant information to formulate research questions or hypotheses	Interpreting the variety of information found for formulating research questions or hypotheses
Investigation	Analyzing the collected data to answer the research questions or control the hypotheses	Planning the procedure of data collection to gather data necessary for answering the research questions or checking the hypotheses	Interpreting the variety of procedures (their pros and cons) that could be applied in data collection
Conclusion	Analyzing the quality of the inquiry process and how well the conclusions enable to answer the research questions or check the hypotheses (or what are the related limitations)	Planning the procedure of evaluating the quality and limitations of the conclusions	Interpreting the findings in order to draw a conclusion to answer the research question or to accept or reject the stated hypotheses; to understand the scope of the conclusions
Discussion	Analyzing what and how to present to others and how to incorporate and understand the feedback of others	Planning the discussion with others and the presentation of procedure or outcomes of any of the inquiry phases to others	Interpreting the meaning of the outcomes to others so that they understand the main point and can give feedback; interpreting the feedback of others and reflecting on the learning process

11 years of age; an average of 10 years) in 146 schools in Estonia. Fifty percent of the students were boys (N = 933) and fifty percent were girls. The test was completed in the Estonian language (the language of instruction, although it was not the mother tongue for all the students) by 1639 students (88%) and in the Russian language (they also studied science at school in Russian) by 229 students (12%). Data from 92 students who completed the test in 2019 were used for analyzing the predictive validity of the test. These students were from a school where the SIT-PE test results were comparable to the national average. Most of the deviations of all four outcomes on all levels were below 10%. The two exceptions were the following: compared to the national average, there were 10.3% less students with a zero level of planning skills and 16.2% less students with a medium level in interpretation skills.

The procedure of conducting tests

The SIT-PE test is conducted in an electronic environment—the Examination Information System (EIS) developed in Estonia by Foundation Innove (later reorganized as a new institution) and used for most of the state level tests in Estonia. However, it can also be used in an international context—the SIT-PE test has already been made available for administration in English and Greek. The EIS enables the administrators to enter tasks and compile tests from the existing tasks. There are many options to design the tasks; eight different types of questions were used in the final version of the SIT-PE test: (1) multiple-choice questions with only one correct option, (2) multiple-choice questions with more than one correct option (usually two out of five had to be selected), (3) tasks of pairing up pictures and texts/fragments of text, (4) tasks of forming a sequence of phases, (5) tasks of drawing a graph or pointing at something in the picture, and (6) open-ended questions. The initial version of the test also included tasks with the benefits of the online testing environment, e.g., tasks where students had to use an online simulation to collect data in an experiment, tasks where students had to watch videos, or tasks where they had to find information on the internet in order to evaluate its trustworthiness. Unfortunately, these tasks were left out from the final test due to quality issues revealed in our empirical study.

The EIS also enables the registration of users who could take the tests. In our study, the students were first registered by their schools and then logged in to the electronic environment where they answered all the questions sequentially. It was not possible to move back to the previous questions, because correct answers had to be displayed to students at certain points before moving on in order to test all the expected learning outcomes. However, the students had the possibility to log into the system after their open-ended answers had been scored by the teachers and their levels of outcomes had been calculated for each dimension of the SIT-PE test. They were given not only the results but also guidelines on what to focus on in future learning.

The time limit for taking the SIT-PE test was 60 min. Most of the students completed all tasks in 45 to 60 min. In the first main study (conducted in the second development phase of the test), we also asked the students if they had had enough time for completing the test. This question was answered by 1680 students: 54% reported that there was enough time; 38% said that some time was even left over; and only 8% found that there was not enough time. It should be noted that the test was also taken by language immersion classes or students who study in a school with Estonian as the language of instruction but whose mother tongue is a different language. Thus, we understand that the test might be challenging for some students, but indeed it was not difficult for majority of the students according to their own feedback.

Data analysis

One-parameter Item Response Theory (1PL IRT, i.e., the Rasch measurement model) analysis was applied to evaluate and improve the quality of the SIT-PE test (in the rest of the text we use it so, although it is equivalent to the Rasch analysis). Winsteps® Rasch measurement software was used (Linacre, 2020). We preferred the 1PL IRT analysis to the two-parameter model (2PL IRT), since 1PL IRT analysis allows evaluation of the quality of the items and test by a more rigorous criterion (Fox & Bond, 2015). The 1PL IRT requires that each item has good discrimination and that all items have the same discrimination index (equal to 1), while the 2PL model does not impose such a requirement (e.g., it allows items to vary in terms of their discrimination). The 1PL IRT is especially suitable for developing an instrument, because it helps to identify items that do not meet rigorous criteria of good items. In addition, it enables to identify potential revision that might improve the quality of items in order to meet the requirements of a good item. This is the reason why we chose to use 1PL IRT analysis instead of 2PL IRT. Similarly, 1PL IRT/Rasch analysis has been applied in developing several other tests in science education (see DeBoer et al., 2014; Kuo et al., 2015; Schiefer et al., 2019). The 1PL IRT analysis was used in both the second and third phase of the study. We only introduce results from the third phase, because these show the quality of the final SIT-PE test. In the second phase, the same analyses were performed to improve the test items and to select the items for the third phase. In both cases, we used the partial credit model with an average item difficulty about 0.5 logit higher than students' average ability, which indicates good test targeting.

In order to answer the first research question—i.e., assess the quality of the SIT-PE items—several analyses were made. First, we analyzed how well the items were able to discriminate respondents. For that, an estimation of the discrimination score was used. Although the Winsteps analysis is based on the one-parameter IRT measurement model fixing discrimination of all items at value 1, it still provides a post hoc estimation of item discrimination that can be used as an indication of items that might have an issue in discriminating respondents. This estimation of the discrimination measure was used to identify items with problematic discrimination.

Second, we described the item fit of the test items. The item fit in the test was measured by three indices: infit, outfit, and item-total correlation. Based on these indices, it is possible to evaluate the quality of each item and its potential to contribute to a good measurement of science learning skills. We used a threshold with the following values: for infit and outfit indices, acceptable values are those in the range of .7 to 1.3; for the item-total correlation we take .20 as an acceptable value and .30 as good item-total correlation, as suggested in the Winsteps manual (see http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm).

Third, the variation of the difficulty measure of the test items was also analyzed. The variation of the difficulty of the items was evaluated by the distribution of items in terms of their difficulty, with an average item difficulty set to 0. We assumed that, for a good measurement tool, it would be important to have items that are placed around the middle of the scale, but also items that are significantly simpler or more difficult. In addition, item reliability and student reliability were used to estimate the replicability of item difficulty measures (in case they are used with another sample of participants from the same population) and replicability of measures of student skills (in case another sample of items from the same population of items is used with the same sample of students).

Finally, the quality of the scoring key of each item was analyzed. It was evaluated based on the average score of participants who responded differently on the same item. If the scoring key is clearly defined, then participants who provided an answer that is evaluated as correct should have a higher average total score than students who provided a partially correct answer or a wrong answer. The same is expected when comparing students who provided a partially correct answer and those who provided a wrong answer.

The second research question about the latent variables that can be differentiated with the SIT-PE test was answered using confirmatory factor analysis (CFA). First, the normed chi-square index was calculated. Based on Kline (1998) and Ullman (2001), the model was considered acceptable if the value of the index was below 3 and good if below 2. The model was considered acceptable if the fit indices were the following (see Bowen & Guo, 2011): root mean square error of approximation (RMSEA) ≤ .05, comparative fit index (CFI) ≥ .95, and Tucker-Lewis index (TLI) ≥ .95. In addition, the weighted root mean square (WRMR) was used, as suggested by Yu (2002), in case some of the items are dichotomous, as in the case of the SIT-PE test. The value of the WRMR index should be close to 1.0.

The third research question about the predictive validity of the SIT-PE test was calculated using correlation analysis. The average score of student levels in the four outcomes measured by the SIT-PE test was correlated with the criterion variable, which was the science grade given by their teacher for the same period when the SIT-PE test was conducted. For calculating predictive validity, a subset of the new state level test conducted in 2019 was used. The sample for this subset consisted of 92 students.

IRT analysis was conducted using WINSTEPS 4.0.1., and CFA was done using Mplus Version 7.4 (Muthén & Muthén, 2016) based on the MPlus guidelines (Muthén & Muthén, 1998-2011).

Results

Description of the SIT-PE test and descriptive data of students' scores

We have developed and tested three versions of the SIT-PE test. The items of the first two versions have been further used for compiling the third version of the SIT-PE test as we selected the items that are of good quality and enable measuring, on three levels, students' analytical skills, planning skills, interpretation skills, and science knowledge. Next, we present the findings that will be discussed later in this paper in order to develop the test further. The third version of the test can be used for assessing the learning outcomes of each general inquiry phase described in the framework of Pedaste, Mäeots, et al. (2015), as introduced in Table 2. However, it needs to be noted that this test does not assess all skills in relation to all inquiry phases. For example, in the Orientation phase, only analytical skills are assessed (see Table 2). Science knowledge assessed with the SIT-PE test is not usually linked to any specific inquiry phase, and similar questions could be asked in different phases.

The first version of the SIT-PE test was based on the identification of nine specific learning outcomes that operationalize different inquiry phases. For example, there were two specific learning outcomes in the Orientation phase: understanding and creating scientific texts, and identifying and formulating a problem in a situation.

Table 2 Learning outcomes assessed with the SIT-PE test

Inquiry phase	Outcomes	Items
Orientation	Students understand scientific text and create a simple scientific text. Students identify a problem in a situation and formulate it clearly.	Students read a text and answer questions (analytical skill, basic, or medium level) or create a sentence (analytical skill, high level). Students read a text and formulate a problem that is described there (analytical skill, high level).
Conceptualization	Students formulate research questions and hypotheses.	Students are provided with a problem description and have to formulate or select from a list a research question or hypothesis that contains all elements of a correct question or hypothesis (planning skill, medium or high level).
Investigation	Students design an experiment for collecting data, select appropriate tools and materials, and conduct the experiment.	Students read a research question or hypothesis and list all tools and materials needed in an experiment for answering the question or testing the hypothesis (planning skill, basic or medium level); or plan a sequence of activities needed in this experiment (planning skill, high level); or conduct the experiment using web-based simulations and analyze the collected data (analytical skill, basic or medium level).
Conclusion	Students use or create models in explaining phenomena, processes, or systems. Students solve science-related problems occurring in every-day life and make decision based on scientific knowledge, skills, and values. Students analyze and interpret scientific information and subsequentially draw conclusions and make decisions.	Students select an appropriate model or figure (analytical skill, medium level); or improve a partly prepared model or figure (analytical skill, high level). Students synthesize different textual and visual information with data and make decisions based on more than one type of arguments. (interpretation skill, high level) Students describe data in a table or figure (analytical skill, basic level) and draw conclusions based on these (analytical skill, medium or high level).
Discussion	Students explain and analyze objects, phenomena, and processes, and the cause-effect relationships between them. Students correctly use scientific concepts, symbols, and units.	Students combine information that is presented in the task with their own knowledge (interpretation skill, high level). Students fill in a blank in a text with an appropriate scientific concept (science knowledge, basic, medium or high level).

All items in the SIT-PE test required students to read a text and answer questions, create sentences, or formulate problems. The final test consisted of seven tasks with 24 items in total. Five out of the seven tasks focused on different skills in the context of a complete inquiry process, and two tasks focused only on interpretation skills in the context of decision-making. Each task started with a description of a problem situation around a scientific topic, e.g., how to grow plants or understand the changes in the states of matter. The task continued with several items that are each designed to assess one particular skill or related knowledge. There were 8 items for assessing analytical skills, 5 items for assessing planning skills, 6 items for assessing interpretation skills, and 5 items for assessing science knowledge (see Table 3). There is at least one item for assessing every skill or science knowledge at each of the three levels. The exact test items used in the test are made available to the readers on request if the reader agrees to keep them confidential. Confidentiality is important, because the same items will also be used in the future in the state level science tests in Estonia and in some international comparisons. Indeed, the test can be administered in international studies in collaboration with the authors of the test. It is possible to use it by writing to the first author of the article, and then a confidentiality contract will be signed with the national institution holding the rights of the test. This kind of procedure of giving only limited access to computerized tests is common practice worldwide.

Table 3 Overview of SIT-PE test items

Dimension	Number of items on each level		
	Basic level	Medium level	High level
Analytical skills	3	2	3
Planning skills	1	3	1
Interpretation skills	1	2	3
Science knowledge	2	2	1

An example of a task of the SIT-PE test is provided in the Additional file 1: Example of a test task. In this example, two boys wonder why their pulse level is different when doing cycling training in different terrains. They need to understand what the problem is in the situation and select the two most appropriate answers out of five options (task 1, an item for assessing analytical skills on basic level). The two correct ones are marked in a green color. After that, they move on and cannot come back to change their answers. In the next step, they can read the text about the situation again, but one of the correct problem statements is already given to them in the task. This is presented as one possible correct answer, and there is no direct indication on the correctness of the answer given by the student in order to avoid a decrease in test-taking motivation. In the next step, they need to specify the most correct research question (task 2, an item for assessing planning skills on high level). Again, they need to move on and cannot come back, because in the third step, they are provided with the correct research question and they need to plan an experiment to answer that question. In this example, they are asked what is required for this experiment (task 3, an item for assessing planning skills on medium level). They are not guided to list different categories of requirements; the correct answer (in green) and scoring guidelines (in red) show that the planning skill is good if the students are able to name the required objects from three different categories: subjects (e.g., humans), objects (e.g., bicycles, different terrains), and measurement tools (e.g., sport watch for measuring heart rate). In some other tasks, they might be asked to sequence the steps needed to complete an experiment (an item for assessing planning skills on high level) or to explain why something needs to be done in order to increase the validity of the experiment (an item for assessing planning skills on high level). After planning the experiment, the student moves on again and cannot come back. The fourth step in this example is analyzing the results they have from an experiment. In some other tasks, the students are asked to collect data using virtual web-based tools. In the example, the student has a table and needs to answer two questions based on it. First, the student fills in the gaps in a text (these short answers are scored automatically) (task 4.1, an item for assessing analytical skills on basic level); next, they need to interpret the importance of conducting several trials in the experiment (task 4.2, an item for assessing interpretation skills on medium level). Finally, they proceed to the fifth step, where they have to draw conclusions based on the provided data (task 5, an item for assessing analytical skills on high level). Again, they have to select two best options out of five; however, in some other tasks, they might be asked to answer an open-ended question to draw a conclusion (an item for assessing analytical skills on high level).

Depending on the item, students can get zero points for an incorrect answer, one or two points for a partially correct answer, and three points for a fully correct answer. Students are assigned the highest level of skill or knowledge if they have at least 50% of the maximum score of the items on this level. For example, a student is deemed to be on a high level of analytical skills if they score at least 50% for the items on high level, even if they score, for example, only 40% on a medium level. This is because items on higher levels can only be correctly answered if lower-level skills are present. This way, we also decrease the effect of the contextuality of the items—nothing bad happens when the student is not familiar with all the different contexts of the tasks. The focus of the assessment is on general inquiry skills—mainly on analytical skills, planning skills, and interpretation skills—and slightly less on science knowledge related to the tasks. In addition, a fourth level can be distinguished, a "zero level," which is assigned to the student if none of the item levels (basic, medium, or high) receive a score of at least 50%.

The test results of the sample that took the final SIT-PE test are presented in Fig. 2. The analysis revealed quite a high diversity in students' levels in all four dimensions. The diversity was especially high in analytical skills and interpretation skills and slightly lower in science knowledge, where about two thirds of the students were on high level. This shows that the test is good for differentiating students.

What is the quality of each individual item of the SIT-PE test and their potential to measure science learning outcomes?

We have analyzed the data by 1PL IRT in order to evaluate the quality of each individual item based on item infit and outfit, item-total correlation, and estimated item discrimination as well as to evaluate the potential of the set of items to measure individual differences among primary education students in terms of science learning outcomes. The aim of the study was to combine items into four measures: analytical skills (items starting with "An"), planning skills (items starting with "Pl"), interpretation skills (items starting with "In"), and science knowledge (items starting with "Kn").

The key findings about the quality of each of the 24 items are presented in Table 4. Based on several indices (infit and outfit, item-total correlation, and estimated item discrimination), we can see that 18 items have all indices in the expected range for good and satisfactory items, while 6 items have some indices outside the expected range for some items. In addition, a supplement table (Additional file 2: Data) gives a detailed overview of each item option for all items in the test. First, it presents all data codes and respective scores; e.g., in An1_k1, full credit (3 points) was given to answer code 2 (the student selected both correct options, B and D, out of five options). Partial credit (2 points) was given to students who selected only one of the correct alternatives (only B or only D). Furthermore, supplement (Additional file 2: Data) also shows how many respondents there were in each answer category, their percentage of the sample, and the average ability of students whose answers were classified in the given category. In the case of a good item, if some answer category (code) is scored higher, the average ability of students providing this answer is supposed to be higher than the average ability of students providing an answer that is scored lower.

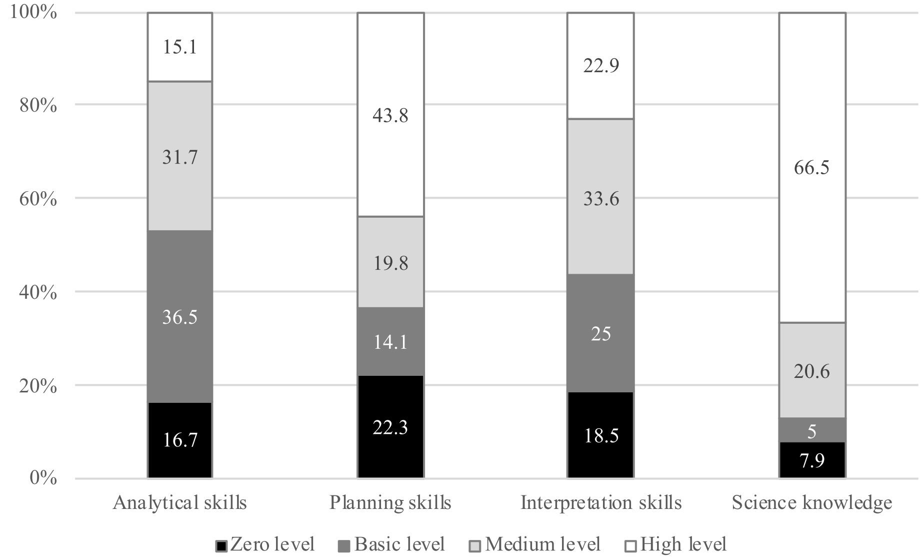


Fig. 2 The ratio of students on different levels of dimensions of the SIT-PE test. In addition, students' perception was used to evaluate the appropriateness of the test for the target group: 10.9% of students found that it was very difficult, 28.6% found it rather difficult, 51.6% moderately difficult, 7.1% rather easy, and 1.7% easy. While 59.2% of the students considered their computer skills definitely good enough for completing the test, 36.2% regarded them as more or less sufficient, and only 4.4% as insufficient

Item An1_s4 has somewhat higher infit and outfit values and a somewhat lower estimated item discrimination. Since this is a partial credit item and the easiest item in the test (the item difficulty is − 1.20), it might be the result of the fact that some high scorers had difficulties solving this item. However, since the item has a good item-total correlation, we can conclude that it is a good item despite the difficulties mentioned above, and that it is useful for differentiating students at the lower level of science inquiry competence. Item An2_v3 has an item-total correlation lower than .20, but it has good infit and outfit indices and good estimated item discrimination. This is a multiple-choice item with 5 alternatives. Based on the analysis of the average ability of students who choose different distractors (see Additional file 2: Data), it can be seen that students who select a correct alternative have just a slightly higher average score than students who select a wrong alternative E. Since other indices are good, we have decided to keep this item for further analysis, but to revise the formulations of alternatives provided to students. Item Kn2_v1 has somewhat higher infit and outfit indices, while other parameters are good. This is a rather difficult partial credit item (.39), so higher infit and outfit indices might suggest a need for a more precise definition of criteria for partial credit and full credit answers. Items An3_k5 and An3_s5 have somewhat higher infit indices and An3_s5 also has a higher outfit index, but since they have a high item-total correlation and a good estimated discrimination index, we have decided to keep them in the analysis. Finally, item Pl2_s2 has somewhat lower infit and outfit indices (that might be the result of a somewhat higher correlation between this item and item Pl2_k3), but since its other indices are quite good, we have concluded that it should be kept in the test. Based on these findings, we can conclude that the SIT-PE test consists of a good set of items, although some of the items can be improved in future.

Since the scoring key is an important part of item quality, we have also analyzed the quality of the scoring key for each item. The evaluation of the item scoring key was based on an analysis of the average ability of students who choose different alternatives in multiple-choice items or students who get different scores on open-ended items. The analysis showed that on all items, students who scored higher on a particular item also had a higher average total score compared with students who provided an incorrect answer or a partially correct answer (see Additional file 2: Data). Only in the multiple-choice item An2_v3, we identified a need for a revision of different alternatives in order to ensure a stronger differentiation between students who select the correct option and students who select an incorrect option. One possibility seemed to be to differentiate a correct and partially correct answer instead of a mere correct and incorrect one. The correct option explained the specific relationship between the variables (the higher the temperature, the faster the evaporation). One of the incorrect options stated that the speed of evaporation depends on temperature. Although it is not as specific as the option describing the exact relationship based on provided data, it is indeed correct. It is

Table 4 Indices of the quality of SIT-PE items evaluated based on the 1PL IRT analysis

Item[a]	Type[b]	Measure[c]	Model SE[d]	Mean infit[e]	Mean outfit[e]	Corr.[f]	Exp. Corr.[g]	Estim. Discrim.[h]
An1_k1	5	− 0.70	0.04	1.07	1.08	.28	.40	0.82
An1_k6	4	− 0.97	0.05	0.96	0.94	.37	.31	1.17
An1_s4	5	− 1.20	0.04	**1.47**	**1.42**	.51	.40	**0.43**
An2_s1	5	0.09	0.04	0.94	0.94	.31	.39	1.06
An2_v3	1	0.31	0.05	1.07	1.13	**.18**	.29	0.77
An3_k5	2	0.62	0.03	**1.36**	**1.38**	.49	.47	1.13
An3_k7	5	0.67	0.04	0.99	1.04	.25	.36	0.94
An3_s5	2	1.29	0.04	**1.34**	0.96	.48	.38	1.17
Pl1_k2	4	− 0.37	0.05	1.00	0.99	.32	.31	1.03
Pl2_k3	2	− 0.23	0.03	0.75	0.81	.47	.52	0.74
Pl2_k4	5	− 0.08	0.04	0.93	0.94	.33	.39	1.07
Pl2_s2	2	0.01	0.03	**0.56**	**0.56**	.60	.51	1.23
Pl3_p2	3	0.50	0.04	0.96	0.96	.42	.38	1.06
In1_r3	5	− 0.74	0.05	1.05	1.04	.52	.40	0.98
In2_m1	5	0.04	0.03	0.96	0.95	.50	.39	1.12
In2_s3	5	0.29	0.05	0.71	0.73	.40	.38	1.40
In3_m2	2	0.87	0.03	0.58	0.65	.51	.44	0.82
In3_r1	1	0.39	0.04	0.94	0.93	.36	.27	1.18
In3_r2	2	− 0.06	0.04	1.12	1.15	.44	.52	1.36
Kn1_t1	4	− 0.89	0.04	0.89	0.87	.45	.31	1.45
Kn1_v2	2	0.66	0.04	0.91	0.91	.47	.47	0.68
Kn2_t2	4	− 0.99	0.03	1.02	1.03	.28	.31	0.91
Kn2_v1	2	0.39	0.05	**1.71**	**1.81**	.38	.49	0.75
Kn3_p1	5	0.11	0.03	0.90	0.91	.34	.39	1.12

Values that are outside of accepted values are marked in bold

[a] Item names are codes where the first two letters show the dimension assessed (An, analytical skills; Pl, planning skills; In, interpretation skills; Kn, knowledge), the number next to them shows the level of the skill assessed (1, basic; 2, medium; 3, high), the letter next to the underscore indicates the task, and the number at the end is the number of the item in the task

[b] Item type: 1, multiple-choice question with only one correct option (radio button); 2, open-ended question coded with 2 or more points (partial credit item); 3, forming a sequence of phases (multiple choice); 4, open-ended question coded dichotomously (correct or incorrect); 5, multiple-choice question with more than one correct option (check boxes, number of correct ones has been defined, e.g., 2 out of 5; selection from a number of pictures)

[c] Item difficulty measure

[d] Standard error of the item difficulty measure

[e] Mean infit and mean outfit refer to Infit MNSQ and Outfit MNSQ indices, which suggests how well students' scores estimated based on item difficulty and students' ability fit to real student scores (items with infit and outfit indices in the range of .7–1.3 are considered good items)

[f] Correlation between item score and respondent ability score estimated based on the 1PL IRT model (items with a correlation higher than .20 are considered satisfactory items; items with a correlation higher than .30 are considered good items)

[g] Expected correlation between each item score and respondent ability score

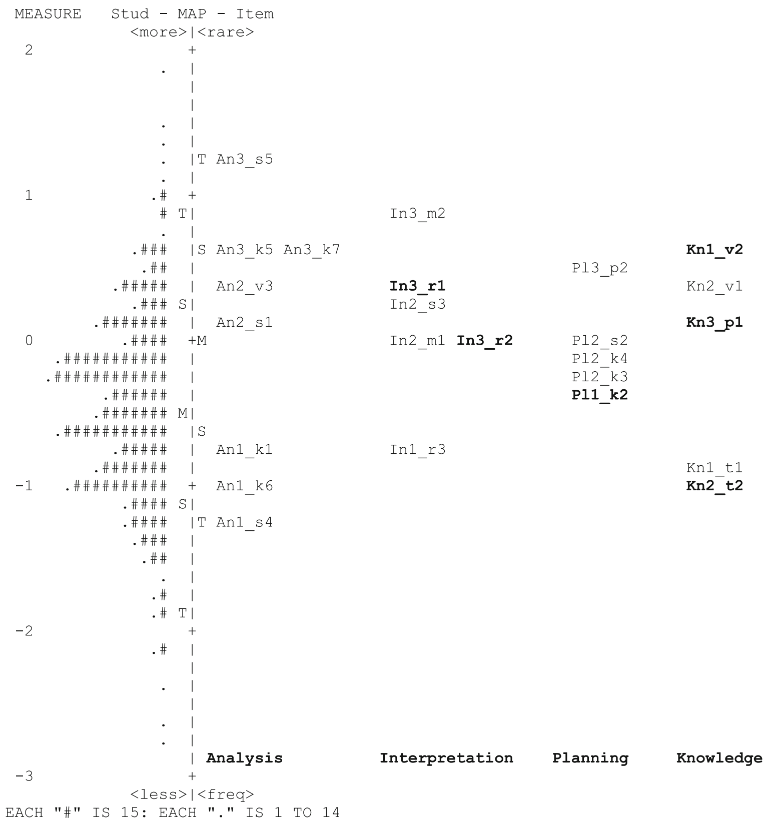
[h] Estimated item discrimination, i.e., what would be the item discrimination index if the data were analyzed by a 2PL IRT model (items with estimated discrimination in the range of .5–2.0 are considered items with satisfactory discrimination)

definitely more correct than the other, incorrect options. Therefore, it might be reasonable to revise the scoring key of this item to differentiate a correct and partially correct answer.

Furthermore, we have analyzed the quality of the SIT-PE test consisting of these 24 items in order to evaluate its quality of measuring science inquiry competence. Global infit and outfit indices were very good for both item difficulty and student ability—for item difficulty, global infit and outfit for 24 items were close to 1 (both have the value 1.01); the values were similar for the measurement of student abilities (global infit 1.02 and global outfit 1.01). The reliability of the estimated item difficulties was 1.0, thus suggesting that it is very likely that we will get the same item difficulties with another, similar sample of students; however, the reliability of estimated student abilities was .79, which is satisfactory. Therefore, the current version of the SIT-PE test allows researchers a reliable measurement of science inquiry competence at the ISCED 1 level of education.

Having concluded that this version of the SIT-PE test can be used for measuring the science inquiry competence of students at the ISCED 1 level of education, we also analyzed the WrightMap to find out the extent to which the current set of 24 items allows a good measurement of different levels of student abilities, as well as whether item difficulties fit our expectations about the level of the competences they are intended to measure (see Fig. 3). At the top of the scale are more difficult items—it is rare that students answer these correctly. At the bottom are the easiest items. On the left, it is shown how many respondents there are on different difficulty levels—at the top are those with the highest competence, and at the bottom are those with the lowest competence (each # equals to 15 students). Based on these data, we can conclude that the current set of 24 items is somewhat more difficult than the average science inquiry competence of students involved in the study (the average item difficulty is 0.47 units higher than the average student ability). The distribution of item difficulties ranges from − 1.20 to 1.29, which means that these items cover various levels of science inquiry competence. More items are placed at the upper part of the scale, meaning that these items enable researchers to differentiate somewhat students with higher science inquiry competence better. Still, it is worth noting that there are enough items enabling researchers to differentiate students who are at the lower part of the scale.

However, the WrightMap also indicated that some items were relatively more or less difficult than we had expected. For example, items In3_r1, In3_r2, Kn3_p1, and Kn2_t2 were easier than we expected, taking into consideration that they are supposed to measure the third or second level of the competence; however, our findings suggest that they are somewhere lower down the scale. In contrast, items Pl1_k2 and Kn1_v2 were somewhat more difficult compared to our intentions. These findings suggest that these items should be checked further in order to get a better understanding as to why they turn out to be relatively more or less difficult. After, they should be revised in the future. Analysis of the distribution of item difficulties according to measured outcomes allows us to conclude that the item difficulties of items measuring analytical skills are distributed according to our expectations. For planning and interpretation skills, the item difficulties of most



```
 MEASURE   Stud - MAP - Item
             <more>|<rare>
  2                +
                .  |
                   |
                   |
                .  |
                .  |
                .  |T An3_s5
                .  |
  1            .#  +
                # T|                    In3_m2
                .  |
             .###  |S An3_k5 An3_k7                                    Kn1_v2
              .##  |                                       Pl3_p2
           .#####  |  An2_v3            In3_r1                         Kn2_v1
             .### S|                    In2_s3
         .#######  |  An2_s1                                           Kn3_p1
  0         .####  +M                   In2_m1 In3_r2      Pl2_s2
     .###########  |                                       Pl2_k4
    .############  |                                       Pl2_k3
          .######  |                                       Pl1_k2
         .####### M|
     .###########  |S
           .#####  |  An1_k1            In1_r3
         .#######  |                                                   Kn1_t1
 -1   .##########  +  An1_k6                                           Kn2_t2
            .#### S|
            .####  |T An1_s4
             .###  |
              .##  |
                .  |
               .#  |
               .# T|
 -2                +
               .#  |
                   |
                .  |
                   |
                .  |
                .  |
                   | Analysis          Interpretation    Planning     Knowledge
 -3                +
             <less>|<freq>
EACH "#" IS 15: EACH "." IS 1 TO 14
```

Fig. 3 The WrightMap presenting difficulty of SIT-PE items and distribution of student abilities. M on the left side of the vertical scale—average ability of students participating in the study. M on the right side of the vertical scale—average difficulty of items included in the study. S on the left side of the vertical scale—one standard deviation from the average ability of students. S on the right side of the vertical scale—one standard deviation from the average difficulty of items. T on the left side of the vertical scale—two standard deviations from the average ability of students. T on the right side of the vertical scale—two standard deviations from the average difficulty of items

items are distributed according to our intentions, while the item difficulties of knowledge items are somewhat different in most cases compared to our intentions; the reasons for this need to be discussed further.

What are the latent variables that can be differentiated with the SIT-PE test?

The second research question focused on testing if the four learning outcomes (analytical skills, planning skills, interpretation skills, and science knowledge) tested with the SIT-PE test could be differentiated empirically as latent variables. Confirmatory factor analysis (CFA) was used to test the factor structure of the test. CFA with all 24 items of the SIT-PE test was with almost acceptable fit indices (χ^2/df = 2.53, RMSEA = .029, CFI = .930, TLI = .922, WRMR = 1.188). However, the latent factors were strongly correlated with each other (from .661 to .898). Next, the correlations were allowed between the residual variances of a few items based on the modification indices and analysis of content of items: An1_k6 WITH An1_s4, because these were exactly the same questions of reading a scale in two different inquiry tasks; In1_r3 WITH In3_r2, because one of the questions explained the reason why the correct answer of the other question was correct; In2_s3 WITH Pl2_k4, because both of them were about the control of conditions in an experiment; and Pl2_s2 WITH Ka2_k3, because both of them were about listing everything that is needed in the experiments. Then, the model fit was improved (see Table 5), but the correlations between latent variables were even higher (from .671 to .924). Therefore, we hypothesized that the model structure could be either unidimensional, second order, or bifactorial. We tested all these models, allowing the same correlations between the residuals of the items as in the case of the correlated factors model. The results are presented in Table 4. It appeared that the unidimensional and bifactorial models are with a significantly worse fit to the data, but the second-order model is with about the same fit indices as the correlated factors model.

The diagrams of the two models with the best fit indices are presented in Fig. 4. Both models have an issue that should be reflected and investigated in the next study. If the correlated factors model has overly high correlations between some of the latent variables then, in the case of the second-order factor model, the first order factor of analytical skills has the value of unexplained variance below zero and the loading of the second-order factor on this first order latent variable is greater than 1. We believe that the correlated model is more promising for future research. The main reasons for taking the correlated model as a starting point for the further development of the model and instrument are related to the fact that the correlated model has somewhat better fit parameters and does not include any negative residual variance. Moreover, a high correlation between four latent dimensions might be the result of the fact that these outcomes are still not differentiated at the early stage of science learning.

Table 5 Fit indices of the confirmatory factor analysis models describing SIT-PE test

Factor model	χ^2	df	χ^2/df	RMSEA	CFI	TLI	WRMR
Correlated factors	509.019	242	2.21	.024	.951	.944	1.071
Unidimensional	686.142	248	2.77	.031	.919	.910	1.257
Second order	515.584	244	2.11	.024	.950	.943	1.080
Bifactor	1152.571	225	5.12	.047	.828	.789	1.699

Correlations are allowed between the following items: An1_k6 WITH An1_s4, In1_r3 WITH In3_r2, In2_s3 WITH Pl2_k4, and Pl2_s2 WITH Ka2_k3

What is the predictive validity of the SIT-PE test?

The predictive validity of the test was measured using correlation analysis to correlate students' results in the SIT-PE test and their science grades in school. Data from 92 students were used in this analysis. First, the data was checked for normality, which revealed that the distribution of data was statistically significantly different from normal distribution according to the Kolmogorov-Smirnov test (p < .01). Therefore, the Spearman rank order correlation was used. The correlation between the students' grades and average scores of the SIT-PE outcomes was .447 (p < .01). The correlations between the grade and different outcomes were even stronger: for analytical skills .620, planning skills .518, interpretation skills .706, and science knowledge .454. All correlations were statistically significant at the .01 level (2-tailed). These moderate or strong correlations show good predictive validity of the SIT-PE test.

Discussion

The aim of the current study was to develop a science inquiry test for primary education level based on the inquiry-based learning framework developed by Pedaste, Mäeots, et al. (2015). According to our knowledge, several self-report instruments for science learning outcomes are available (e.g., Chang et al., 2011), but there are no tests for measuring science competence according to the inquiry-based learning approach at the primary education level. Therefore, we welcomed the result which showed that the items of the SIT-PE test developed in this study were mostly of good quality and that the test also demonstrated a multidimensional structure of science outcomes and good predictive validity. What is also worth noting is that the test meets the strict requirement of item discrimination being equal to 1. In comparison, Scalise and Clarke-Midura (2018) found in developing their measures for capturing inquiry-oriented performance an average item discrimination of .56 with

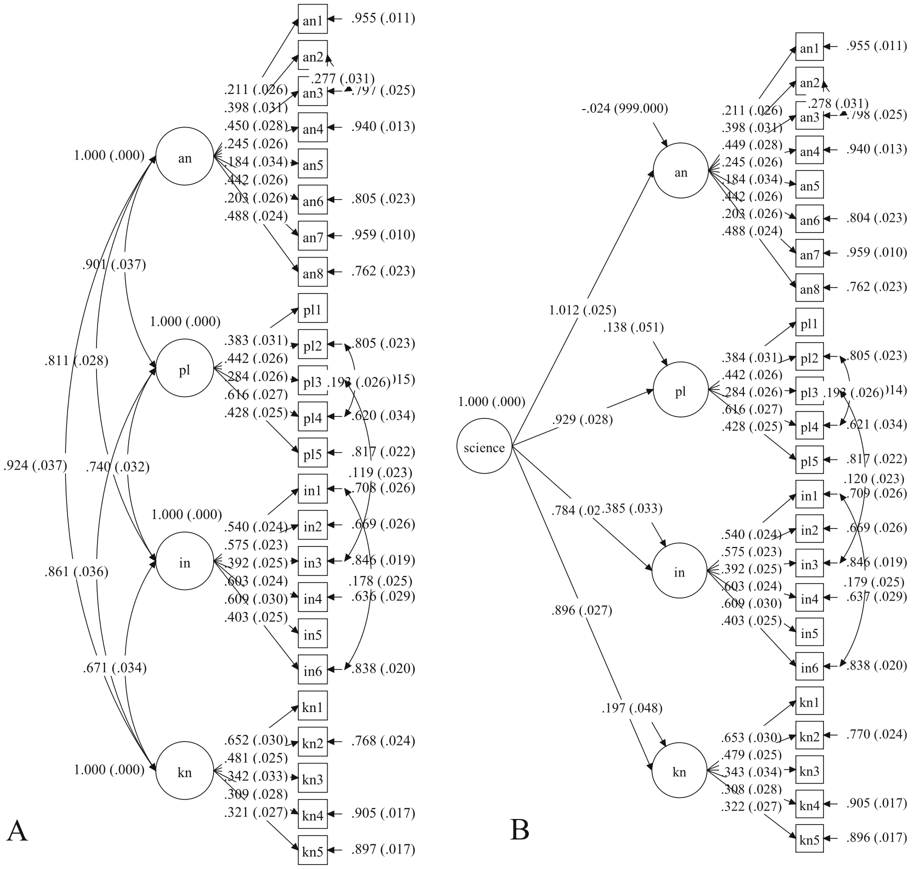


Fig. 4 Correlated factor model (**a**) and second-order factor model (**b**) describing the science inquiry. Correlated factor model (**a**) describing the four factors of analytical skills (An), planning skills (Pl), interpretation skills (In), and science knowledge (Kn) and second-order factor model (**b**) describing the general science competence factor (science) and first order factors describing analytical skills (An), planning skills (Pl), interpretation skills (In), and science knowledge (Kn). The model was created using the data of 24 items of the SIT-PE test (*n* = 1868)

a range from .27 to .72, while it is suggested that discrimination scores between .5 and 2.0 are good for performance tests. However, there are some open issues with respect to each finding that need to be discussed in more detail.

Quality of items in the SIT-PE test

First, we focused on item difficulty. Part of the value of our study was that we used IRT models in the three phases where large data sets were collected. This enabled us to select the best items from one phase to another and to test the stability of item difficulty on different samples of students who participated in the three phases. Furthermore, it enabled us to identify three levels of each science learning skill that were consistent throughout different phases. This has not been the case in developing other known tests for assessing science learning outcomes in the context of inquiry-based learning. For example, Scalise and Clarke-Midura (2018) used the WrightMap to discuss the difficulty of items but did not specify the levels of difficulty. Nevertheless, the levels are very important, because this makes it possible to describe each level and to give feedback and suggestions for further learning based on the description of a slightly higher level of difficulty in the student's zone of proximal development. This was also the case in the SIT-PE test—the students were provided with feedback about their levels in four learning outcomes and suggestions for further learning in each of these dimensions. This supports learners in improving their self-regulation skills and teachers in becoming more aware of the diversity of their students in order to plan further studies and provide formative feedback.

The test items for assessing planning skills were all of good quality. In interpretation skills, there was an issue with item In3_r2, which was easier than had been expected based on previous data collection. The item difficulty did not seem to derive from the interpretation skills necessary for interpreting different data in making a decision, but from the fact that students needed to select two alternatives to get full credit on this item (option B and option E). Actually, most of the students selected only option B without recognizing that option E should also be selected. Moreover, when we compare students who chose only option E or only option B (and got partial credit), it seems that option E required a more advanced level of interpretation skills in order to be recognized as a correct option (see data related to item In3_r2 presented in Additional file 2: Data). The fact that the two correct options were on different levels of complexity, and that students can get partial credit regardless of the option they select, might be an explanation as to why this item turned out to be easier than we had expected. Thus, in the revision of the items, it should be ensured that if two options need to be selected, the answers should be more or less equally appropriate.

In analytical skills, one item was found where the item infit and outfit indices were slightly higher than expected. This means that the students who got a good total score had a slightly higher rate of incorrect answers on this item and the ones with a lower total score answered it considerably well. In total, 54% of the students answered it correctly, 21% partially correctly, and 25% incorrectly. The item was about reading the scale of a thermometer. The correct answer was 24 degrees, and the partially correct answer allowed them a ± 1°. However, the scale had markings only in every 5°, and the degrees were marked in numbers only in every 10°. Actually, the precise value was not important in the inquiry. What was important was to keep the value on the same level in all experiments, but it did not matter if it was 24, 25, or even 20 or 30°. The conclusions of the experiment would have been the same. Therefore, one explanation for the high infit and outfit indices might be that students with better results in science pay less attention to unimportant aspects, whereas those who focus more on details might not understand the bigger picture behind the details; therefore, the latter might fail more often in the assessment of generic skills, which were the focus of the SIT-PE test. In the future, the test items should be improved so that the details are asked only where it is important to focus on the details.

In addition, there was one item assessing analytical skills where the correlation with the total test score was slightly lower than .20, which we consider a minimal threshold. It also indicates a potential issue with infit and outfit indices, but these were very well in the boundaries of the suggested values. In this item, students had to analyze a data table and select the most appropriate description of the relation between two variables introduced in the table. Two options out of five were marked with almost the same frequency (33% of the students chose the correct one and 35% the incorrect one), and their correlation with the total score of the test was almost the same (indeed slightly higher for the correct one). This suggested that the wrong option should be analyzed in order to identify why it might be misleading and attract students with high analytical skills. By a detailed analysis of the two options, we came to the conclusion that both options might be considered correct, although one of them (the one considered correct in the scoring key) is more precise. The correct option states "The higher the temperature, the faster the water evaporates", whereas the incorrect one says, "The rate of evaporation of the water depends on the temperature." The incorrect option does not specify how one variable depends on the other. The item could be improved when the scoring key is revised by differentiating correct and partially correct options. The same principle could be applied in the case of other items as well to further improve the quality of the items.

In assessing science knowledge, more significant findings emerged that need to be discussed. Two items appeared to be simpler and one more difficult than expected based on the previous data collection. The explanation for this phenomenon might be that the knowledge gained in science depends very much on contextual factors, e.g., what is the focus in one or another school, and what is meaningful to one or another student. Therefore, we concluded that the science knowledge levels could only be specified for items if the item difficulty is first determined in every new study. This seems to be also the reason why the students' level of science knowledge was mostly high, and why the test did not provide much variability (see Fig. 1). The highest level of knowledge was assessed with only one item, which was expected to be with high difficulty (based on previous data collections) but was easier than one knowledge item on basic level and one on medium level. By analyzing the items, we proposed a hypothesis for further studies—there are different types of knowledge that should be analyzed on different scales. Some of the knowledge items focused on specific scientific terms (e.g., the student had to know the three characteristics of clean water: *transparent, flavorless,* and *odorless*; however, these are terms that we do not need to use in our everyday discussions) and some others on generic knowledge (in this case simpler) that is often used in everyday life in different contexts (e.g., that the weight of seeds could be described in *grams* or plants need *light* in order to grow). A similar issue might be the reason why one item was the easiest one, although it was on a medium level based on previous studies. In this item, students had to finish a sentence about why the sun is needed to grow plants. The correct answers were either *light* or *energy*. Light could be considered a generic everyday life term—everybody can see that light reaches plants. However, energy is a much more abstract concept and requires relevant scientific knowledge. In everyday experience, one does not see that energy reaches from the sun to the plants. Scientifically, it would be more correct to treat only the answer related to the concept of energy as a correct answer or, if both answers were to be considered correct, the answer related to energy should get full credit and the answer related to light should get partial credit.

Another issue that requires attention in developing the test further is designing items assessing knowledge that are more related to items assessing skills. In the current test, these were only connected through the topic, not more. For example, in an inquiry task on the effect of temperature on the evaporation of water, the students could make an inference based on a table of data collected in an experiment without relying on the knowledge that was tested in the same task. The two items assessing knowledge in the same task required students to explain what is melting and how to characterize clean water. This knowledge was not necessary to perform well in the inquiry task. Therefore, the correlations between the inquiry item and the two knowledge items were only .043 and .077. For the inquiry task, much more generic knowledge was necessary, e.g., knowing that the value of something could be read from the intersection of a column and row in a table.

Latent variables differentiated with the SIT-PE test

Despite the minor issues associated with some of the quality indicators of a few items in the SIT-PE test, we found that the test could be used for assessing science learning outcomes developed on an inquiry framework. However, we also showed that the science learning outcomes assessed with the SIT-PE test have a multidimensional structure. In this respect, our findings support those of Scalise and Clarke-Midura (2018), who showed that at least two latent traits could be differentiated in assessing students' inquiry-oriented performance in science. In our study, the fit indices of the correlated factors model and second-order factor model showed that analytical skills, planning skills, interpretation skills, and science knowledge were separate latent factors, and the unidimensional or bifactorial models had worse fit indices. Thus, whereas Scalise and Clarke-Midura (2018) differentiated only two latent variables, SIT-PE enabled us to distinguish four factors. The two factors identified by Scalise and Clarke-Midura (2018) were described as inquiry and explanation. In our study, it appeared that the four dimensions correlate strongly. One interpretation of this might be that there is still a general factor present when describing general science competence above all specific competences; however, this model was not revealed in our study, because the assessment of different skills and knowledge was not sufficiently related. The final version of the SIT-PE test was compiled by selecting the required items from the best tasks developed in the previous studies—items of good quality that were necessary for assessing four dimensions on three levels. It turned out that some topics were used only for assessing some particular dimensions. One task assessed only knowledge with two items, another two assessed interpretation with two and three items, one task assessed knowledge with one item and analytical skills with two items, one assessed knowledge with one item and planning with another item, one assessed planning with three items and analysis with four items, and only one task assessed three dimensions out of four: planning and interpretation with one item and analysis with three items. Therefore, it would be good to design new test items in the future so that a significant number of test items in different dimensions would be on the same topic. Then, the bifactorial and second-order factor model could be tested again. Another solution for improving the model might be revising the theoretical model. For example, Scalise and Clarke-Midura (2018) had a correlation of only .42 between the latent variables in their two-dimensional model, so we might consider revising our model to distinguish fewer latent variables. In this case, we might see that our findings are more or less in line with those of Scalise and Clarke-Midura (2018). In our case, the lowest average correlation with other latent variables was in the case of interpretation skills, which are the closest to one of the latent variables—explaining (analyzing the results of the inquiry, drawing conclusions, and communicating results)—introduced by Scalise and Clarke-Midura (2018). Our latent variables, describing analytical skills and planning skills, could be more related to their inquiry variable, which included posing questions, designing investigations, and carrying out the investigation. However, our science knowledge variable could not be associated with one of the two variables in the study of Scalise and Clarke-Midura (2018). In addition, the correlations in our study were still significantly higher between all latent variables.

Another explanation for the high correlations between the latent variables in the factor model might be the age group of our respondents. Tucker-Drob (2009) has described the phenomenon that in younger age groups, some skills could be merged more easily into general skills, and when students get older, the different dimensions become more distinct. In order to test this hypothesis, the SIT-PE test could also be used with older students; in Estonia, state level science tests are conducted every 3 years, and a similar science test is administered at the beginning of the seventh grade (age 13 to 14) to assess the science outcomes achieved after the second level of primary school. However, when we compare our results with the study of Kuo et al. (2015), then the factors in the SIT-PE test are even better distinguished from each other, although the sample of Kuo et al. was older—secondary school students from the 8th and 11th grade. Kuo et al. (2015) differentiated four inquiry abilities but showed that the correlations between the factors ranged from .87 between questioning and analyzing to even .96 between experimenting and explaining.

Predictive validity of the SIT-PE test

Finally, the predictive validity of the SIT-PE test revealed that regardless of the various possibilities for improving the quality of some of the test items, the test is already in a state where it could be used for assessing science learning outcomes in the context of inquiry-based learning. Regarding predictive validity, it was interesting that the criterion measure—science grades given by the teachers in school—was more strongly correlated with different skills than with the science knowledge or average level of different learning outcomes assessed with the test. All these correlations were remarkably higher than the correlations usually reported in employment tests, where predictive validity has often been assessed. The strongest correlation in the SIT-PE test was between students' grades and interpretation skills. This might show that teachers place great value on interpretation and other science skills in grading students, which might also explain why Estonian students do well in the international PISA test (OECD, 2019); however, this needs to be studied further, because in the current study, the predictive validity was calculated based on data collected only in one school. Indeed, this school represented an average school in Estonia according to the SIT-PE test results. In addition, it should be noted that whereas grading in school takes place in high-stakes settings, the SIT-PE test was a low-stakes test for the students. Therefore, the correlation between the grades and SIT-PE test results might be even higher if the SIT-PE test were to be used for grading students. Currently, this was not the aim—the test was administered to give students and teachers feedback about different learning outcomes of the inquiry process and to provide specific recommendations for next learning steps.

Conclusions and implications

We designed a science test for assessing students' analytical skills, planning skills, interpretation skills, and science knowledge at primary education level. IRT analysis showed that the test as a whole is of good quality, as are most of the test items and the scoring key of the items. The assessment of knowledge appeared to be slightly more complicated and requires further inquiry. We hypothesized that several types of knowledge should be differentiated, e.g., specific scientific knowledge and general transferable inquiry knowledge. In addition, we admitted that the knowledge assessed in the SIT-PE test was not really needed for the inquiry activities. Thus, the test is already good for assessing science skills but could be developed further to provide more insight into different types of knowledge.

We also found that confirmatory factor analysis almost equally supports both the correlated factors model and the second-order factor model in describing the science competence assessed by the SIT-PE test. Both models confirmed the multidimensional structure of the science learning outcomes: analytical skills, planning skills, interpretation skills, and science knowledge. Indeed, it appeared that the correlations between the latent variables were rather strong. In this context, it is also important to further explore how much the second-order factor—general science competence—is described by science skills, and how much is explained by science knowledge. The relationship between science skills and knowledge is also worth further investigation.

Despite the open questions regarding further development of the SIT-PE test, the study showed that the test has a good predictive validity in its current form. Therefore, we can suggest its use in international studies for assessing primary school students' learning outcomes in the context of inquiry-based learning. However, there are also some limitations that need to be taken into account when applying the test. First, the quality of the test was good, but some levels of certain dimensions were assessed only with one item. Therefore, the test could be even further improved by developing more items in these particular categories. Second, the test has been implemented on large samples and more than once, but only in the context of Estonia, even though it has been administered in the Russian language as well in addition to Estonian. Therefore, the test versions translated and adapted to other languages (e.g., English and Greek to date) need to go through psychometric analyses again using the same analysis methods as in the current study. In addition, the applicability of the items for assessing science knowledge needs to be checked in different contexts, because the knowledge might depend on the curriculum of a particular country. This might apply to content knowledge, in particular, but this should not be an issue in assessing inquiry knowledge. The latter was not assessed with the current version of the SIT-PE test and could be a potential improvement for the next version of the test. Third, the stability of the difficulty of the items for assessing science knowledge needs to be studied further to avoid the situation where the items describe different levels every time there is a new sample. Fourth, the time needed to complete the test is currently 45 to 60 min, which might be quite demanding for 10- to 11-year-old students. Although only a minority of the students reported that they did not have enough time to complete the test, it might be that fatigue had some effect on the validity of the data gathered with the last items of the test. Therefore, it needs to be tested further if some items could be left out from the test without significantly lowering the quality of the test. For example, in some cases, the level of a particular inquiry outcome could be assessed with fewer items if the items have very similar psychometric properties in repeated tests and if they have a higher maximum score to ensure variance among students in order to differentiate them with the test. However, despite its limitations, the test is already applicable as a state level science test in other countries than Estonia or as a test for assessing students' science skills in scientific experiments. The current set of items proved to be very good. This means that we can measure science inquiry outcomes—skills and knowledge—at the primary school level of education. In addition, we also learned which items cause difficulties for students—this is important knowledge for further studies.

One more area of improvement in the SIT-PE test for further research is the design and analysis of new test items, which would include the benefits of online testing, e.g., interactive tasks where students have to conduct experiments using simulations. In our development process, these types of tasks were left out from the final version of the SIT-PE test owing to their unsatisfactory quality; however, in the future, more items could be tested to find the ones with good quality.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40594-021-00278-z.

Additional file 1. Example of a test task.
Additional file 2. Data.

Abbreviations

CFA: Confirmatory factor analysis; IEA TIMSS test: International Association for the Evaluation of Educational Achievement Trends in International Mathematics and Science Study; IRT: Item Response Theory; OECD PISA: Organization for Economic Co-operation and Development, Program for International Student Assessment; SIT-PE: Science Inquiry Test for Primary Education

Acknowledgements

The authors thank all schools who carried out the tests locally, teachers and other experts who contributed in developing the tests, and staff of Education and Youth Authority who provided technical and other types of support in preparing and conducting the tests.

Authors' contributions

MP developed the test items, conceptualized the study, designed the methodology, executed analysis of data, and wrote the manuscript as the main author. AB contributed in conceptualizing the study, designing the methodology, interpreting the results, and writing the manuscript. ER contributed in developing the test items, collected data, and contributed in interpreting the results. The authors read and approved the final manuscript.

Authors' information

MP is a Professor of Educational Technology at the Institute of Education of the University of Tartu, Estonia.
AB is a Professor of Developmental and Educational Psychology and the Head of Institute of Psychology at the University of Belgrade, Serbia. He is also a Visiting Professor at the Institute of Education, University of Tartu, Estonia.
ER is a Science Expert of Education and Youth Authority, Estonia.

Funding

This research has been partly funded by the Estonian Ministry of Education and Research and European Social Fund and Estonian Research Council institutional research grant IUT34-6.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare no competing interests.

Author details

[1]University of Tartu, Salme 1a, 50103 Tartu, Estonia. [2]University of Belgrade, Cika Ljubina 18-20, Belgrade 11000, Serbia. [3]Education and Youth Authority, Lõõtsa 4, 11415 Tallinn, Estonia.

Received: 26 June 2020 Accepted: 16 February 2021
Published online: 17 March 2021

References

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*, 1–18. https://doi.org/10.1037/a0021017.

Archer-Kuhn, B., Lee, Y., Hewson, J., & Burns, V. (2020). Growing together: Cultivating inquiry-based learning in social work education. *Social Work Education*, 1–21. https://doi.org/10.1080/02615479.2020.1839407.

Areepattamannil, S., Cairns, D., & Dickson, M. (2020). Teacher-directed versus inquiry-based science instruction: Investigating links to adolescent students' science dispositions across 66 countries. *Journal of Science Teacher Education*. https://doi.org/10.1080/1046560X.2020.1753309.

Arini, M. D., Suratno, & Yushardi (2019). Analysis pattern of student communication skills in science process in inquiry learning: Study of case study learning in regional schools Jember Coffee Plantation. In *Journal of Physics: Conf. Series*, (p. 1211). https://doi.org/10.1088/1742-6596/1211/1/012104.

Bowen, N. K., & Guo, S. (2011). *Structural equation modelling*. New York: Oxford University Press.

Chang, H.-P., Chen, C.-C., Guo, G.-J., Cheng, Y.-J., Lin, C.-Y., & Jen, T.-H. (2011). The development of a compe- tence scale for learning science: Inquiry and communication. *International Journal of Science and Mathematics Education*, *9*(5), 1213–1233. https://doi.org/10.1007/s10763-010-9256-x.

Constantinou, C. P., Tsivitanidou, O. E., & Rybska, E. (2018). What is inquiry-based science teaching and learning? In O. E. Tsivitanidou, P. Gray, E. Rybska, L. Louca, & C. P. Constantinou (Eds.), *Professional development for inquiry-based science teaching and learning*, (pp. 1–23). Cham: Springer. https://doi.org/10.1007/978-3-319-91406-0_1.

De Jong, F., Kollöffel, B., van der Meijden, H., Kleine Staarman, J., & Janssen, J. J. H. M. (2005). Regulative processes in individual, 3D and computer supported cooperative learning contexts. *Computers in Human Behavior*, *21*(4), 645–670. https://doi.org/10.1016/j.chb.2004.10.023.

De Jong, T., Sotiriou, S., & Gillet, D. (2014). Innovations in STEM education: The Go-Lab federation of online labs. *Smart Learning Environments*, *1*(1), 3. https://doi.org/10.1186/s40561-014-0003-6.

De Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*, 179–202. https://doi.org/10.3102/00346543068002179.

DeBoer, G. E., Quellmalz, E. S., Davenport, J. L., Timms, M. J., Herrmann-Abell, C. F., Buckley, B. C., et al. (2014). Comparing three online testing modalities: Using static, active, and interactive online testing modalities to assess middle school students' understanding of fundamental ideas and use of inquiry skills related to ecosystems. *Journal of Research in Science Teaching*, *51*(4), 523–554. https://doi.org/10.1002/tea.21145.

Fox, C. M., & Bond, T. G. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*, (3rd ed.,). New York: Routledge.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching. *Review of Educational Research*, *82*, 300–329. https://doi.org/10.3102/0034654312457206.

Jones, L. R., Wheeler, G., & Centurino, V. (2015). TIMSS 2015 science framework (Ch. 2). In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks*. Chestnut Hill: Routledge https://timssandpirls.bc.edu/timss2015/downloads/T15_FW_Chap2.pdf. Accessed 26 June 2020.

Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, *40*, 898–921. https://doi.org/10.1002/Tea.10115.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Kuo, C. Y., Wu, H.-K., Jen, T. H., & Hsu, Y. S. (2015). Development and validation of a multimedia-based assessment of scientific inquiry abilities. *International Journal of Science Education*, *37*(14), 2326–2357. https://doi.org/10.1080/09500693.2015.1078521.

Kuter, S., & Özer, B. (2020). Student teachers' experiences of constructivism in a theoretical course built on inquiry-based learning. *Eğitimde Nitel Araştırmalar Dergisi – Journal of Qualitative Research in Education*, *8*(1), 135–155. https://doi.org/10.14689/issn.2148-2624.1.8c.1s.7m.

Linacre, J. M. (2020). *Winsteps1 (Version 4.5.4) [Computer Software]*. Beaverton: Winsteps.com http://www.winsteps.com. Accessed 26 June 2020.

Liu, C., Zowghi, D., Kearney, M., & Bano, M. (2021). Inquiry-based mobile learning in secondary school science education: A systematic review. *Journal of Computer Assisted Learning*, *37*, 1–23. https://doi.org/10.1111/jcal.12505.

Mäkitalo-Siegl, K., Kohnle, C., & Fischer, F. (2011). Computer-supported collaborative inquiry learning and classroom scripts: effects on help-seeking processes and learning outcomes. *Learning and Instruction*, *21*(2), 257–266. https://doi.org/10.1016/j.learninstruc.2010.07.001.

Misra, D. (2020). Using inquiry-based learning in executive education programmes. *Journal of Workplace Learning*, *32*(8), 599–613. https://doi.org/10.1108/JWL-12-2019-0149.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide*, (6th ed.,). Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2016). Mplus. Version 7.4 [Computer software]. Los Angeles, CA: Muthén & Muthén.

Mullis, I. V. S., & Martin, M. O. (Eds.) (2017). *TIMSS 2019 assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center http://timssandpirls.bc.edu/timss2019/frameworks. Accessed 26 June 2020.

National Research Council (2000). *Inquiry and the national science education standards*. Washington, DC: National Academy Press.

OECD (2013). PISA 2015: Draft science framework. http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Science%20Framework%20.pdf. Accessed 26 June 2020.

OECD (2019). *PISA 2018 Results (Volume I): What students know and can do*. Paris: PISA, OECD Publishing. https://doi.org/10.1787/5f07c754-en.

Oguz, A. A., & Aybars, T. (2019). Using the inquiry-based learning approach to enhance student innovativeness: a conceptual model. *Teaching in Higher Education*, *24*(7), 895–909. https://doi.org/10.1080/13562517.2018.1516636.

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections*. London: Nuffield Foundation.

Pedaste, M., Mäeots, M., Leijen, Ä., & Sarapuu, S. (2012). Improving students' inquiry skills through reflection and self-regulation scaffolds. *Technology, Instruction, Cognition and Learning*, *9*(1–2), 81–95.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., Van Riesen, S. A., Kamp, E. T., … Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003.

Pedaste, M., Siiman, L., De Vries, B., Burget, M., Jaakkola, T., Bardone, E., … Veermans, K. (2015). Ark of inquiry: Responsible research and innovation through computer-based inquiry learning. In H. Ogata, W. Chen, S. C. Kong, & F. Qiu (Eds.), *Proceedings of the 23rd international conference on computers in education*. Ishikawa: Asia-Pacific Society for Computers in Education.

Pöntinen, S., Kärkkäinen, S., Pihlainen, K., & Räty-Záborszky, S. (2019). Pupil-generated questions in a collaborative open inquiry. *Education Sciences*, *9*(2), 1–15. https://doi.org/10.3390/educsci9020156.

Scalise, K., & Clarke-Midura, J. (2018). The many faces of scientific inquiry: Effectively measuring what students do and not only what they say. *Journal of Research in Science Teaching*, *55*(10), 1469–1496. https://doi.org/10.1002/tea.21464.

Schiefer, J., Golle, J., Tibus, M., & Oschatz, K. (2019). Scientific reasoning in elementary school children: Assessment of the inquiry cycle. *Journal of Advanced Academics*, *30*(2), 144–177. https://doi.org/10.1177/1932202X18825152.

Shanks, R. A., Robertson, C. L., Haygood, C. S., Herdliksa, A. M., Herdliska, H. R., & Lloyd, S. A. (2017). Measuring and advancing experimental design ability in an introductory course without altering existing lab curriculum. *Journal of Microbiology & Biology Education*, *18*(1), 1–8. https://doi.org/10.1128/jmbe.v18i1.1194.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology*, *45*(4), 1097–1118. https://doi.org/10.1037/a0015864.

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick, & L. S. Fidell (Eds.), *Using multivariate statistics*, (4th ed., pp. 653–771). Needham Heights: Allyn & Bacon.

Wu, P., & Wu, H. (2020). Constructing a model of engagement in scientific inquiry: Investigating relationships between inquiry-related curiosity, dimensions of engagement, and inquiry abilities. *Instructional Science*, *48*, 79–113. https://doi.org/10.1007/s11251-020-09503-8.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Unpublished doctoral dissertation)*. Los Angeles: University of California Los Angeles.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.