UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Assessing students' ability in performing scientific inquiry: Instruments for measuring complex science skills in primary education

Kruit, P.M.; Oostdam, R.J.; Van den Berg, E.; Schuitema, J.A.

DOI
10.1080/02635143.2017.1421530

Publication date
2018

Document Version
Final published version

Published in
Research in Science and Technological Education

Link to publication

Citation for published version (APA):
Kruit, P. M., Oostdam, R. J., Van den Berg, E., & Schuitema, J. A. (2018). Assessing students' ability in performing scientific inquiry: Instruments for measuring complex science skills in primary education. *Research in Science and Technological Education*, *36*(4), 413-439. https://doi.org/10.1080/02635143.2017.1421530

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:05 Sep 2026

RESEARCH IN SCIENCE & TECHNOLOGICAL EDUCATION
2018, VOL. 36, NO. 4, 413–439
https://doi.org/10.1080/02635143.2017.1421530

Routledge
Taylor & Francis Group

Check for updates

Assessing students' ability in performing scientific inquiry: instruments for measuring science skills in primary education

Patricia M. Kruit[a], Ron J. Oostdam[a,b], Ed van den Berg[a] and Jaap A. Schuitema[b]

[a]Centre for Applied Research in Education, Amsterdam University of Applied Sciences, Amsterdam, Netherlands; [b]Faculty of Social and Behavioural Sciences, Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands

ABSTRACT

Background: With the increased attention on the implementation of inquiry activities in primary science classrooms, a growing interest has emerged in assessing students' science skills. Research has thus far been concerned with the limitations and advantages of different test formats to assess students' science skills.

Purpose: This study explores the construction of different instruments for measuring science skills by categorizing items systematically on three subskill levels (science-specific, thinking, metacognition) as well as on different steps of the empirical cycle.

Sample: The study included 128 fifth and sixth grade students from seven primary schools in the Netherlands.

Design and method: Seven measures were used: a paper-and-pencil test (PPT), three performance assessments, two metacognitive self-report tests, and a test used as an indication of general cognitive ability.

Results: Reliabilities of all tests indicate sufficient internal consistency. Positive correlations between the PPT and the three performance assessments show that the different tests measure a common core of similar skills thus providing evidence for convergent validity. Results also show that students' ability to perform scientific inquiry is significantly related to general cognitive ability. No relationship was found between the measure of general metacognitive ability and either the PPT or the three performance assessments. By contrast, the metacognitive self-report test constructed to obtain information about the application of metacognitive abilities in performing scientific inquiry, shows significant – although small – correlations with two of the performance assessments. Further explorations reveal sufficient scale reliabilities on subskill and step level.

Conclusions: The present study shows that science skills can be measured reliably by categorizing items on subskill and step level. Additional diagnostic information can be obtained by examining mean scores on both subskill and step level. Such measures are not only suitable for assessing students' mastery of science skills but can also provide teachers with diagnostic information to adapt their instructions and foster the learning process of their students.

KEYWORDS

Scientific inquiry; performance assessment; primary education; science skills

CONTACT Patricia M. Kruit p.m.kruit@hva.nl

© 2018 Informa UK Limited, trading as Taylor & Francis Group

Introduction

With the increased attention toward the implementation of inquiry activities within primary science classrooms, a growing interest has emerged in assessing students' science skills, which are the skills involved in generating and validating knowledge through scientific investigations. Traditionally, most tests, whether intended for small- or large-scale assessment, have been paper-and-pencil formats consisting of multiple-choice items and/or open-ended questions (Hofstein and Lunetta 2004). Examples of such tests are the 'Test of Enquiry Skills' by Fraser (1980) and the test for assessing science achievement for the Third Mathematics and Science Study (TIMSS) (Martin et al. 1997).

In line with the increased understanding of how students learn, alternative assessment formats such as the use of performance assessments (Harmon et al. 1997; Shavelson, Baxter, and Pine 1991; National Research Council [NRC] 2012) have been considered. In a performance assessment (PA), students perform small experiments by interacting actively with materials. PAs are regarded as 'investigations that recreate to some extent the conditions under which scientists work and elicit the kind of thinking and reasoning used by scientists when they solve problems' (Shavelson, Solano-Flores, and Ruiz-Primo 1998).

Research has been concerned with the limitations and advantages of the different test formats. A paper-and-pencil test (PPT) can be administered easily, rated reliably and students are familiar with the format (Harlen 1991). The major disadvantages are that a PPT lacks authenticity (Shavelson, Baxter, and Pine 1991) or, in other words, the assessment does not reflect the activities of a real-life inquiry (Davey et al. 2015), and may be influenced considerably by reading ability (Harlen 1991). While PAs are considered more authentic (Ennis 1993; Davey et al. 2015) they are also more cost and labor-intensive to administer. Due to the open format, reliable rating is complicated (Davey et al. 2015) and students are often not familiarized with this test format which may negatively influence test performance (Kind 1999).

Prior research shows that different levels of content knowledge – defined by OECD (2017) as 'knowledge of the facts, concepts, ideas and theories about the natural world that science has established' – may affect test performance as well (Harlen 1991; Roberts and Gott 2006; Eberbach and Crowley 2009). Content knowledge may have more influence on a PA than a PPT because a PA is designed around one science topic of which individual students may or may not possess prior knowledge, while a PPT contains questions about several topics. Different strategies have been used to mitigate for the content dependency of items. For example, the TIMSS 2015 Science Framework (Jones, Wheeler, and Centurino 2013) attempted to minimize the influence of content knowledge by assessing science skills using content where it was necessary to reason with the concepts. The Next Generation Science Standards specifically integrated content knowledge and skills in the goals for the 'practices'. The concept of practices is used to integrate the science process skills and content knowledge to 'emphasize that engaging in scientific investigation requires not only skill but also knowledge that is specific to each practice' (NRC 2012, 30). The Assessment of Performance Unit (APU) controlled for content knowledge by minimizing the amount of content in assessments (Harlen 1986). In the Science Teachers' Action Research (STAR) project in the UK, skills were assessed using multiple items relating back to one theme (Harlen 1991) which is similar to a PA in which all items refer to one particular context.

Research that focuses on measuring science skills in both primary and secondary education has not only shown small correlations between PPTs and PAs (Baxter et al. 1992; Baxter

and Shavelson 1994; Lawrenz, Huffman, and Welch 2001; Roberts and Gott 2006; Hammann et al. 2008) but also between PAs designed to measure the same science skills (Gott and Duggan 2002; Pine et al. 2006). These small correlations between the same and different test formats have not only been attributed to differences in students' content knowledge (Shavelson, Baxter, and Pine 1991; Gott and Duggan 2002), but also to inconsistencies in rating and occasion sampling variability. Occasion sampling variability occurs when students perform the same task differently on different occasions (Ruiz-Primo, Baxter, and Shavelson 1993).

The lack of convergence between different tests intending to measure the same science skills suggests at the same time that underlying cognitive demands may not always be evoked equally (Millar and Driver 1987; Shavelson, Baxter, and Pine 1991; Messick 1994). Research shows that, for example, correlations between measures of science skills and general cognitive ability in grade 9 vary from small (Tamir 1988; Song and Black 1992) to large (Lawson 1989; Baxter et al. 1992) indicating that certain components of science tests may be related more to general cognitive ability and less to skills specific for science (Gott and Duggan 2002; Pine et al. 2006; Roberts and Gott 2006).

A major concern in this regard is that science skills are a rather 'ill-defined domain' (Gobert and Koedinger 2011). Science skills – also referred to with terms such as 'inquiry skills' or 'investigation skills' (Harlen and Qualter 2009) – usually indicate a wide variety of activities related to planning and conducting investigations and interpreting results (Gott and Duggan 1995; Alonzo and Aschbacher 2004; Harlen and Qualter 2009). Abrahams and Reiss (2015) make an additional distinction between process skills such as planning, predicting, and experimenting, and practical skills which are more specific such as handling a microscope.

Science skills are generally defined based on the activities in which scientists engage during authentic research (Lederman and Lederman 2014). In the framework for K-12 science education, scientific inquiry is represented by three domains of activities: investigating, developing explanations and solutions and evaluating data as evidence for the proposed theories and models (NRC 2012). The NRC emphasizes that students should learn about what scientists do when designing and carrying out their own inquiries. However, for assessing science skills, it is necessary to more accurately define the cognitive demands underlying the inquiry activities. In this context, Osborne (2014) argues that part of the problem lies in the focus that educators and teachers put on only the practical aspects of scientific inquiry which applies to both primary (Roth 2014) and secondary education. This limited operationalization of science skills results in ignoring the wide variety of cognitive abilities called upon in scientific investigations. For instance, different abilities are employed when handling a microscope than when identifying patterns in data. Consequently, science skills are frequently assessed using tests that are not systematically constructed and based upon a clear operationalization of cognitive demands which underlie scientists' actual activities. Furthermore, tests often emphasize the practical side of inquiry such as controlling variables. For systematic test construction, different types of the underlying skills need to be distinguished, identified (Sternberg 1985) and systematically included in test designs.

The aim of the present study is to explore to what extent structuring assessments by distinguishing between underlying skills will improve convergence between tests, attain more validity by including all aspects of inquiry, and offer the possibility of obtaining diagnostic information on students' performance. To this end, each activity performed in a

scientific inquiry was classified by determining which of the following skills primarily underlies the activity: science-specific skills, thinking skills, or metacognitive skills.

Science-specific skills refer to the ability to apply procedural and declarative knowledge for correctly setting up and conducting a scientific experiment (Gott and Murphy 1987). These skills can be classified as lower order thinking (Newmann 1990) or reproductive thinking (Maier cited in Lewis and Smith 1993), and are characterized by knowledge recall, comprehension, the routine employment of rules, and simple application (Goodson 2000). Students performing a scientific inquiry are required to recall the facts and rules about how to conduct scientific experiments, such as identifying and controlling for variables, observing and measuring and using simple measurement devices. They must then use and apply this knowledge to – for example – select the appropriate procedures and organize the data into tables (Gott and Murphy 1987; OECD 2017). Science-specific inquiry skills defined as such include the practical skills as discussed by Abrahams and Reiss (2015), but they pertain to cognitive processes as well.

In addition to the above-described science-specific skills, students apply more general thinking skills to make sense of the data and connect the observations to scientific theories (Osborne 2015). Thinking skills include the higher order thinking skills, also frequently referred to as critical thinking (Moseley et al. 2005). A distinction is often made between the philosophical interpretation of critical thinking (evaluating statements and judging) and the interpretation made by psychologists who emphasize the problem-solving aspect. The latter approach is more commonly utilized in scientific inquiry (Lewis and Smith 1993).

Thinking skills involve manipulating information that is in nature complex because it consists of more than one element and has a high level of abstraction (Flavell, Miller, and Miller 1993). Concepts and rules are put together and applied to a new situation. The application of thinking skills involves interpreting, analyzing, evaluating, classifying, and inferring information (Newmann 1990; Moseley et al. 2005). In accordance with Bloom's taxonomy, thinking skills such as analyzing and synthesizing are considered to have higher levels of complexity (Bloom 1956). Many of these thinking skills are abundantly applied in scientific investigations. For example, when making appropriate inferences from different sources of data (Pintrich 2002) or when identifying features and patterns in data, thinking skills will predominantly underlie these particular aspects of a scientific inquiry. Zohar and Dori (2003) even argue that science skills such as formulating hypotheses or drawing conclusions can be classified as higher order thinking skills since they have the same characteristics.

Metacognitive skills are in general considered to be a particular type of higher order thinking skill (see for discussion Lewis and Smith 1993). What distinguishes metacognitive skills from general thinking skills is that they involve active executive control of the mental processes (Goodson 2000) or 'thinking about thinking' (Kuhn 1999; Kuhn and Dean, Jr. 2004). In this study, metacognitive skills refer to self-regulatory skills and include planning, monitoring, and evaluating task performance (Flavell, Miller, and Miller 1993; Schraw and Moshman 1995; Pintrich 2002). Planning refers to the selection of appropriate strategies and the allocation of resources that effect performance. Monitoring refers to one's awareness of comprehension and task performance. For instance, checking to see whether one is still on track during the task. Evaluating refers to the appraisal of the products and regulatory processes of learning, for instance when re-evaluating one's goals and conclusions. Although metacognitive skills are considered to play an important role in many types of cognitive activities (Zohar and Barzilai 2013), these skills influence the quality of the scientific inquiry process,

which in particular demands self-regulation and knowledge and use of metacognitive strategies (Schraw, Crippen, and Hartley 2006). For instance, a student who is aware of the shortcomings of a particular inquiry may be able to improve his or her performance in a subsequent scientific inquiry. To become a scientific thinker, students need to acquire metacognitive skills in order to understand, direct, monitor and evaluate their own higher order reasoning (Kuhn 1989).

To further reduce the lack of convergence between tests, the following main activities ('steps') within the empirical cycle were used as a general blueprint for test construction: (1) formulating a research question, (2) designing an experiment, (3) formulating a hypothesis, (4) measuring and recording data, (5) analyzing data, (6) formulating a conclusion, and (7) evaluating. Although scientists do not move linearly through the three domains of activity (NRC 2012) and merely use it as a reporting device (Kind 1999), the empirical cycle reflects all of the aspects of a scientific inquiry which are included as learning objectives in most curricula. Deploying the empirical cycle as a blueprint for test construction ensures that the same activities of scientific inquiry are included in each test and thus ensures construct validity (Solano-Flores et al. 1999). Furthermore, systematically assembling these activities within tests may provide a useful scaffold, especially for students in primary education who have little inquiry experience (Donovan, Bransford, and Pellegrino 1999; cf. White and Frederiksen 2000).

In summary, we explored the construction of different instruments for measuring science skills in grades 5 and 6 of primary education. In contrast to current measures, we aimed for a systematic construction of instruments by assigning items to the different activities of the empirical cycle and by categorizing them in relation to science-specific skills, thinking skills, and metacognitive skills. In this way, we ensured that tests contained the major aspects of scientific inquiry while doing justice to the different cognitive demands which are often overlooked by teachers when assessing science skills (Osborne 2014). Although the underlying skills are not measured directly because the tests aim only to measure the activities performed in a scientific inquiry, it is still possible to obtain a reflection of students' mastery of science-specific skills as well as thinking and metacognitive skills. In addition, the influence of prior content knowledge is controlled for as much as possible.

Furthermore, we examined to what extent the different instruments measure science skills in relation to general cognitive ability and also whether the categorization of items on underlying skill (subskill) level (science-specific, thinking, and metacognition) and step level of the empirical cycle might provide additional diagnostic information. Hence, the following research questions were addressed:

(1) Can students' ability in performing scientific inquiry be measured in a reliable manner?
(2) To what extent is the measurement of students' ability in performing scientific inquiry related to their general cognitive ability?
(3) Can students' ability in performing scientific inquiry be validly measured by means of different assessment instruments?
(4) To what extent do measurements on subskill and step level provide additional diagnostic information to the overall measurement of students' ability in performing scientific inquiry?

Method

Participants

All measuring instruments were administered to 128 students (55% female, 45% male) with a mean age of 11.4 (SD = .64) from seven primary schools in the Netherlands. Seventy-five students (59%) were in grade 5 and 53 (41%) were in grade 6. Some students had prior experience with scientific investigations because of lessons provided within the regular school science curriculum. Science skills had not previously been assessed by means of a PPT or a PA at these schools.

Measuring instruments

Paper-and-pencil test

Items for the PPT were selected from large-scale assessments and other sources (e.g. SOLpass.org) based on the following criteria. Construct validity was maintained by assigning items to the different steps of the empirical cycle and by categorizing them in relation to the primary subskill which underlies the particular activity performed in the item (Table 1). For instance, one of the items contained a short description of an experiment and the data measured. The students were asked to draw a graph of these data. This item was categorized as a science-specific skill and simultaneously assigned to the step of 'recording and organizing data'. The PPT contained items that measured thinking and science-specific skills (see Figures A1–A4 in Appendix 1 for some example items). Items on metacognition such as choosing alternative strategies or evaluating learning gains were not included because answers are based on self-assessments and cannot simply be scored as correct or incorrect (Shavelson, Carey, and Webb 1990). Limited test time available at schools resulted in balancing time-consuming open-ended questions, such as providing an explanation or making a graph, with more time-efficient multiple choice questions.

To ensure content validity, university lecturers in the fields of biology and physics education assessed all items for correct representation of the phenomena. In addition, these content experts checked that items were correctly classified to subskill and step level. Next, primary school teachers verified the formulation of the items and the content familiarity for grades 5 and 6 students. As a result, minor adjustments were made, such as substituting relatively unfamiliar words with more commonly used words. Finally, a small group of five students of similar age were asked to complete the initial draft of the PPT and to explain their answers in an informal interview to check comprehensibility of content and language. As a result, some items were revised or deleted. For example, one item required students to look at a drawing of a cat and interpret its mood. However, it turned out that only students who owned a cat were able to interpret the cat's behavior as shown in the drawing. This item was therefore deleted.

The preliminary version of the PPT was piloted in two rounds with, respectively, 117 and 158 students from grades 5 and 6. Based on the results of these pilot studies items with item total correlations below .15 were deleted, resulting in a final version with 46 items (Table 1).

A scoring model was developed for assessing answers to the open-ended questions. To ensure scoring validity, possible answers were first formulated by content experts and then fine-tuned based on students' answers. Criteria for awarding points were based on the level of complexity of the answers, meaning that the more elements the answer needed, the more

Table 1. Distribution of multiple choice and open-ended items in the PPT, classified to subskill level.

Item description	Number of items	Multiple choice	Open-ended
Thinking: formulate hypothesis	5	3	2
Thinking: control variables	4	4	–
Thinking: identify features, patterns, contradictions in data	6	3	3
Thinking: make inferences informed by evidence and reason	6	5	1
Thinking: relate conclusion to hypothesis/draw conclusion	7	7	–
Science-specific: formulate research question	6	6	–
Science-specific: observe/measure correctly	6	6	–
Science-specific: organize data	6	2	4
Total number of thinking skills	28		
Total number of science-specific skills	18		

points could be awarded. For instance, drawing a graph involves (a) labeling the axes, (b) putting the data points in the right place, and (c) drawing a line of best-fit (see Appendix 2 for an example of a scoring model).

For administration purposes, the items of the final version of the PPT were divided over two test booklets based on an optimal split half. Each test booklet contained 18 multiple choice and 5 open-ended items. Administration of each booklet in the present study took about 45 min.

Performance assessments

Based on PAs in previous large-scale studies for grades 5 and 6, three tasks were developed with topics suitable for students of this age: *Skateboard*, *Bungee jump*, and *Hot chocolate*. *Skateboard* was based on the PA 'Rolling Down Hill' (Ahlbrand et al. 1993). *Bungee jump* and *Hot chocolate* were based on task formats in TIMSS (Martin et al. 1997).

All three PAs concern comparative investigations: students are asked to examine the relationship between two variables (Shavelson, Solano-Flores, and Ruiz-Primo 1998). In *Skateboard,* students must roll a marble (the 'skateboard') down a ruler (the 'hill') to examine the relationship between the distance of the marble on the ruler (slope) and the distance the marble covers at the end of the ruler while pushing a paper wedge forward. Comparable investigations must be performed in *Bungee jump* (students examine how the length of a rubber band may change by hanging additional weights) and *Hot chocolate* (relationship between the amount of hot water and the rate of cooling).

Each PA is constructed according to the same template following the various activities (steps) of the empirical cycle (Table 2). Subsequently, the different activities are categorized as 'science-specific', 'thinking', or 'metacognitive'. As mentioned before, this categorization is based on the prevailing skill of a particular activity. For instance, the activity of planning an experiment is related to describing the setup of the investigation and the way in which results will be noted. This activity is therefore categorized as science-specific, although it requires thinking and metacognition as well.

To reach a high quality of content validity, the university lecturers in the field of biology and physics education assessed all items regarding clarity of formulation and the main subskill and activity to be measured. According to Clauser (2000), it should be taken into consideration that subject-matter experts may be too focused on details not appropriate for primary school students. Therefore, primary school teachers were also requested to assess all items on the same characteristics. As a result, minor adjustments were made. For example,

Table 2. Blueprint of the performance assessments with items classified to subskill and step level.

Item	Description of activities	Step level	Subskill level	Score
1	Students formulate their own research question	Research question	Science-specific	0-1-2
2	Students design experiment:	Design	Science-specific	
	Description of experimental setup			0-1-2-3
	Description of how results will be noted			0-1-2
3	Students formulate hypothesis	Hypothesis	Thinking	0-1-2
4	Students note their results in a table students make themselves	Measure & record	Science-specific	0-1-2-3
5	Students make a graph of the data they gathered:	Measure & record	Science-specific	
	Axes			0-1-2
	Line graph			0-1-2
6	Students interpret the results by relating two variables	Analyze	Thinking	0-1-2
7	Students extrapolate the results	Analyze	Thinking	0-1-2
8	Students draw a conclusion about relationship	Conclusion	Thinking	0-1-2
9	Students formulate support for their conclusion	Conclusion	Thinking	0-1-2
10	Students relate the hypothesis to the conclusion	Conclusion	Thinking	0-1-2
11	Students identify differences between plan and execution of experiment and explain reason(s) of differences or in absence of differences, give suggestions to improve the experiment	Evaluate	Metacognitive	0-1-2
12	Students give suggestions to extend the experiment	Evaluate	Metacognitive	0-1-2
13	Students draw a conclusion related to the context	Conclusion	Thinking	0-1-2
14	Students formulate their learning gains about inquiry	Evaluate	Metacognitive	0-1-2
	Maximum score			34

in the empirical cycle, formulating a research question is usually followed by formulating a hypothesis and then planning the experiment. However, for primary school students formulating a hypothesis *after* planning the experiment provides students with additional scaffolding.

Preliminary versions of all three PAs were piloted with 70 grades 5 and 6 students. Based on the outcomes several adjustments were made regarding the formulation of instructions, questions and task structure.

Simultaneously, a scoring rubric was developed for each PA. Scoring validity was attained in the following ways (Kane, Crooks, and Cohen 1999). Criteria for awarding points were expressed as detailed descriptions of the elements that should be included in students' answers. University teachers as content experts assessed the criteria for awarding points to the different levels of proficiency of possible answers, meaning that when the answer contains more elements, a higher level of proficiency is reached. Depending on the number of elements more points are awarded (Table 2). In addition, teachers considered whether the criteria were feasible for grades 5 and 6 students. Students' responses obtained from the pilots were used to fine-tune the scoring criteria and examples were added to illustrate the different levels of proficiency (see Appendix 3 for an example of a scoring rubric).

As shown in Table 2, each PA contained 14 quantifiable items to be completed in about 45 min. Scoring of items was based on students' answers which were written down in notebooks. The rationale behind using students' answers is that in authentic inquiry in which all activities are performed, these written responses can be interpreted as a summary of the actual scientific investigation (Kind 1999). An important advantage of using notebooks is that it makes it possible to score and analyze the students' work after the event has occurred (Schilling et al. 1990). Furthermore, scoring based on written answers has proven to be a good alternative for real-time observation (Ruiz-Primo, Baxter, and Shavelson 1993; Solano-Flores et al. 1999) and is assumed to provide a valid indication of the students' potential performance in real-life inquiry (Harmon et al. 1997).

Raters were thoroughly trained to interpret the criteria as intended and to award points to students' answers. During training sessions scoring rubrics were fine-tuned with additional examples of possible answers.

Metacognitive self-report tests

Two metacognitive self-report tests were used. The first test was based on the Junior Metacognitive Awareness Inventory [Jr. MAI], a self-report inventory for grades 3–5 developed by Sperling et al. (2002). Jr. MAI has been used in other research and proven to be a valid measure for metacognition (see Sperling et al. 2002 for discussion). Moreover, Jr. MAI has been validated specifically for measuring metacognition in young students and is therefore appropriate for our purposes. The test is easy to administer and score.

Jr. MAI consists of 12 items with a three-choice response (never, sometimes, or always). Of these 12 items, 6 items evaluate metacognitive knowledge. For example: 'I know when I understand something'. The other 6 items are directed at assessing regulation of cognition. For instance, 'I think about what I need to learn before I start working'. For the present study, the 12 items were translated into Dutch by the researcher, an educational scientist and a primary school teacher. The translations were then translated back into English and compared with the original Jr. MAI of Sperling et al. (2002).

The second metacognitive self-report test – Science Meta Test (SMT) – was designed to measure metacognitive self-regulatory skills, including orientation/planning, monitoring, and evaluation (Schraw and Moshman 1995). In contrast to the more general Jr. MAI, items were constructed specifically to obtain information about the extent to which metacognitive skills are applied in the PAs. For example: 'While doing measurements, I continued to verify that I was following my plan'. Submitting the items to a small sample of students showed that no reading or comprehension problems occurred. The final version of the SMT consisted of 13 items with a three-point scale (not, a little, a lot).

Combined Cito scores

Most primary schools participate in a semi-annual assessment of The National Institute for Educational Testing and Assessment [Stichting Cito Instituut voor Toetsontwikkeling] to monitor students' achievement. Scores are used to advise students for continuing education. Since Cito scores of Reading comprehension and Arithmetic/mathematics significantly correlate with other tests measuring general ability, Cito scores can be considered a valid indication of general cognitive ability (de Jong and Das-Smaal 1995; Bartels et al. 2002; te Nijenhuis et al. 2004). Because scores on these reliable standardized tests were available, a separate cognitive ability test – which would have required additional time and effort for the schools and students – was not administered.

Ability is expressed by different levels which indicate the actual performance level of a student compared to a norm group (A = upper 25% of all children, B = 25% above mean, C = 25% below mean, D = next 15% below C, E = lowest 10%). As a result of the norm-based interpretation, students' test scores can be compared within and between grades. For Reading comprehension and Arithmetic/ mathematics both reliability scores (indicated by Accuracy of Measurement) are high, >.87 and >.95, respectively, for grades 5 and 6 (Janssen et al. 2010; Weekers et al. 2011). For this study, the mid-term tests scores were transformed into a five-point scale (A = 5 to E = 1). A combined Cito score (CCS) was established by summing the scores of both tests.

Administration procedure

Research assistants administered the PPT to all 128 students in a classroom setting and the PAs individually in groups of four to a maximum of eight students. Each research assistant received extensive training and followed detailed protocols for test administration.

Tests were administered to all students on two separate occasions with a time interval of 8–10 weeks. On each occasion tests were administered in the same order: first one split halve of the PPT followed by a performance assessment. *Skateboard* was administered on the first occasion and the other two PAs on the second occasion. To control for sequencing effects, administration of the PAs on the second occasion was randomly rotated. The two metacognitive self-report tests were administered on the second occasion after *Bungee jump*.

Scoring procedure

All handwritten answers to open-ended questions for the PPT and PAs were transcribed to typed text. By doing this, raters were not able to recognize or be influenced by handwriting.

Three raters, all master students, received separate training for the scoring of open-ended questions in the PPT and PAs. Before every training session, raters were provided with the test material, the scoring rubrics and a set of answers of students reflecting various performance levels. Interrater reliability was estimated by determining intraclass correlation (*ICC*, two-way random, absolute agreement) for each rating session on a random sample of an average of 12% of the scores. To avoid bias, raters were instructed to score one item for all students before moving on to the next item. In this way, more sensitivity to different performance levels regarding a particular item was achieved. Depending on the interrater agreement reached, additional discussion of rating differences was initiated. After establishing satisfactory interrater reliability (varying from .71 to .92, single measures ICC) administered tests were randomly distributed to be scored by individual raters. On average, the rating process took 10 min per student for the PPT and 20 min per student for a PA.

Method of analysis

The data-set contained the scores on all measures taken of a total of 128 primary school students. Variables were examined for accuracy of data entry, missing values, and distributions. There were less than 5% missing values on the variables of the metacognitive self-report tests and the Cito tests. Little's MCAR test was not significant indicating that no identifiable pattern exists for the missing data ($\chi^2 = 1596.125$, d$f = 1593$, $p = .47$). EM imputation was performed for missing items of the metacognitive self-report tests. Imputation for missing items of the Cito tests was not possible because only the overall test score was available. All underlying assumptions (i.e. normality) were met.

Overall scores and reliabilities were calculated for all measures. In addition, scores and reliabilities on subskill and step level were calculated for the PPT and PAs. Average item scores were calculated for Jr. MAI and the SMT. Pearson zero-order and partial correlations were calculated to examine discriminant and convergent validity.

Results

Descriptive statistics

Table 3 presents the means and standard deviations for all measures. Both PPT and PAs show normal distributions of scores, indicating that no floor or ceiling effects occur. All PAs show relatively low means indicating a high difficulty level with an average of 31% of the highest score possible. Four students had a score of 0 for *Skateboard* and for *Bungee jump*, and for *Hot chocolate*, only two students had a score of 0. There is therefore no indication of a substantial floor effect. A repeated-measures ANOVA revealed significant differences between the three PAs (Wilks' lambda = .921, $F(2, 126) = 5.40$, $p = .006$, $\eta_p^2 = .079$). Pairwise comparisons demonstrated differences between *Skateboard* and *Bungee jump* (mean difference 1.242, 95% CI [.191, 2.294], $p = .015$) and between *Bungee jump* and *Hot chocolate* (mean difference .977, 95% CI [.102, 1.851], $p = .023$). When interpreting scores as a measure of difficulty, these differences indicate that *Bungee jump* is somewhat easier than the other two PAs. Scores (indicated by average item scores) on Jr. MAI are relatively high, while scores on the SMT are more evenly spread. The mean of CCS shows that students perform around average. For the PPT, *Bungee jump* and *Hot chocolate* girls outperformed boys. However, results show that

Table 3. Means and standard deviations for all measures.

	Max score	Min	Max	*M*	SD
PPT	60	11	49	31.12	8.67
PAs					
Skateboard	34	0	23	10.07	5.56
Bungee jump	34	0	24	11.31	5.09
Hot chocolate	34	0	21	10.34	4.88
Jr. MAI	3	1.50	2.75	2.32	.25
SMT	3	1.23	3.00	2.16	.31
CCS	10	2	10	6.92	2.23

Note: Max score = *maximum score possible of test;* Min = *minimum score of student;* Max = *maximum score of student.*

Table 4. Reliability coefficients of all measures (Cronbach's *α*).

	α
PPT	.82
PAs	
Skateboard	.72
Bungee jump	.67
Hot chocolate	.69
Jr. MAI	.62
SMT	.66

Note: See method section for reliability of the CCS.

the boys had significant higher scores for Jr. MAI. Except for the PPT on which students in grade 6 scored higher than grade 5 students, there were no differences in scores between grades.

Test reliabilities

In order to answer the first research question, test reliabilities were calculated. Cronbach's *α* coefficient for the PPT can be considered good (Table 4). Deleting items would not substantially improve the reliability coefficient. The *α*-coefficients for the three PAs as well as the metacognitive self-report tests indicate sufficient internal consistency.

Relationship between science measures and combined Cito score (CCS)

To examine discriminant validity (research question 2), interdependency between science measures and CCS was explored. Medium to large correlations (Cohen 1988) were found between CCS and respectively the PPT and PAs (Table 5). No significant correlations were found between CCS and the two metacognitive self-report tests.

Relationship between science measures

To find evidence of convergent validity (research question 3), we explored to what extent the overall scores for the different science measures correlate. Because findings show that the sciences measures are related to the combined Cito score (Table 5) correlations were controlled for the scores on this test.

Table 5. Correlations (Pearson's *r*) between science measures and CCS as a reflection of general cognitive ability (d*f* = 121).

	CCS
PPT	.67*
PAs	
Skateboard	.51*
Bungee jump	.55*
Hot chocolate	.42*
Jr. MAI	.04
SMT	−.01

*Correlations are significant at $p < .001$ (2-tailed).

Table 6. Correlations (Pearson's *r*) between science measures controlling for CCS (d*f* = 120).

	1	2	3	4	5
PPT					
PAs					
Skateboard	.43*				
Bungee jump	.34*	.42*			
Hot chocolate	.40*	.46*	.57*		
Jr. MAI	−.11	.04	.04	.05	
SMT	−.02	.12	.20**	.19***	.65*

*Correlations are significant at $p < .001$ (2-tailed); **$p = .029$; ***$p = .038$.

Medium to large significant positive correlations were found between the PPT and all three PAs (Table 6). Tests for comparing dependent correlations measured on the same subjects (Steiger 1980) revealed a significant difference between *Bungee jump* and *Hot chocolate* and between *Bungee jump* and *Skateboard* ($Z = 1.96$, $p = .05$). The average correlation between the three PAs is higher ($r = .48$) than between the PPT and the PAs ($r = .39$), indicating that the PAs tap into somewhat different skills than the PPT. Still, the medium to large correlations between the PPT and the three PAs reinforce that both test formats measure a common core of similar skills, other than that of general cognitive ability alone.

No significant correlations occurred between Jr. MAI and the PAs. These results may indicate that the general metacognitive skills measured by the Jr. MAI were not reflected by the PAs. An alternative explanation may be that the Jr. MAI lacked the sensitivity to measure metacognitive skills.

The SMT, designed specifically to obtain information about the extent to which metacognitive skills are applied in the PAs, correlated significantly with *Bungee jump* and *Hot chocolate* but not with *Skateboard*. Although correlations are small, this might indicate that the metacognitive skills measured in the PAs may be of a more task-specific nature than those obtained by measuring general metacognitive skills.

Additional diagnostic information on subskill and step level

Descriptive statistics

In order to answer research question 4, we explored whether an analysis on subskill and step level provided additional diagnostic information about students' performance levels. Each item in the assessments was assigned to the subskill that most underlaid the concerning

Table 7. Means and standard deviations of standardized scores (0–10) on subskill and step level.

	PPT		PA		Total	
	M	SD	*M*	SD	*M*	SD
Subskill						
Thinking	5.63	1.52	3.00	1.53	4.14	1.33
Science	4.67	1.73	3.26	1.59	3.83	1.49
Meta	–	–	2.99	1.54	2.99	1.54
Step empirical cycle						
Research question	7.73	2.28	4.71	2.79	6.22	2.12
Design experiment	3.95	3.24	2.76	1.60	3.01	1.53
Hypothesis	6.22	2.20	4.60	2.52	5.47	1.90
Measure and record	3.84	1.84	3.21	1.98	3.53	1.69
Analyze	6.21	1.64	2.76	2.02	4.62	1.51
Conclusion	4.88	2.13	2.73	1.59	3.21	1.43
Evaluation	–	–	2.99	1.54	2.99	1.54

Table 8. Scale reliabilities of items measuring the same subskill per tests and aggregated for the PPT, PAs, and PAs together (Cronbach's α).

	Thinking	Science	Meta	Total
PPT	.73	.69	–	.82
PAs				
Skateboard	.59	.55	.36	.72
Bungee jump	.58	.31	.37	.67
Hot chocolate	.53	.54	.36	.69
Aggregated scores of PAs	.77	.74	.64	.86

Note: *Aggregated scores* represent scores for all three performance assessments as one construct.

activity. Then, scores of each scale were obtained, which reflected the main underlying skill applied to that particular cluster of activities. To explore how students performed in the particular aspects of the empirical cycle, scores per step level were also calculated.

In Table 7, the descriptive statistics on subskill and step level are given for all measures. To facilitate comparison between tests, scores were converted to standardized scales between 0 and 10. Means and standard deviations show that on subskill level scores are somewhat low but similar within tests. Scores on step level are more evenly spread within tests, indicating differences between steps in terms of difficulty. Examining the scores for the PPT shows, for instance, that designing an experiment (3.95) seems to be more difficult than formulating a hypothesis (6.22). In general, scores on step level are higher for the PPT than for the PAs, suggesting that different test formats elicit the same skills but are applied in different ways (see also Table 3). For instance, formulating a research question is not the same as identifying a research question among different multiple-choice options.

Reliabilities on subskill and step level

In Table 8, scale reliabilities on subskill level are presented for the PPT and the PAs. Because the three PAs are similar in respect to format, wording, number of items and structure, separate scores could be combined to obtain a reliable aggregated score. By doing so, variance caused by task effects was reduced. Internal consistency specified by Cronbach's α indicates coherent scales on subskill level.

In addition to presenting Cronbach's α for scales on subskill level (Table 8), also the internal consistency on step level was investigated. Table 9 presents scale reliabilities on step level

Table 9. Reliability of the paper-and-pencil test (PPT) and the performance assessments (PA) on empirical step level indicated by Cronbach α.

	PPT		PA	
	Number of items	α	Number of items	α
Research question	6	.53	3	.54
Design experiment	4	.58	3	.56
Hypothesis	5	.26	3	.44
Measure and record	12	.62	6	.59
Analyze	12	.56	6	.47
Conclusion	7	.37	12	.66
Evaluation	–	–	9	.64

Table 10. Correlations (Pearson's *r*) between scores on subskill level and CCS (d*f* = 121).

	CCS (general cognitive ability)
PPT thinking	.59*
PPT science	.61*
PA thinking	.51*
PA science	.55*
PA metacognitive	.27**

*Correlations are significant at $p < .001$ (2-tailed); **$p = .003$.

for the PPT and the PAs. The items of the PAs assigned to metacognition represent the evaluation step as well. Cronbach's α coefficients are in general weak to moderate, indicating that ability scores on the level of the steps should be interpreted with caution.

Relationship with combined Cito score (CCS)

To investigate to what extent general cognitive ability influences ability on subskill level, correlations were calculated for each subskill. Large correlations were found between CCS and the subskills thinking and science-specific (Table 10). For metacognition, the correlation with CCS is small, indicating that items in which metacognitive skills are called upon may tap less into general cognitive ability, reflected by CCS, than do thinking and science-specific items. However, this result might also be explained by the small number of items used to measure metacognitive skills. Alternatively, a study by Veenman, Wilhelm, and Beishuizen (2004) showed that in this age group, students' metacognitive skills are better able to predict learning performance than their intellectual ability. This was attributed to the fact that the tasks in the experiment were too complex for the students. The small correlation between the metacognitive items and the CCS may be attributable to a similar effect.

Table 11 shows significant correlations between CCS and scale scores on step level. In general, correlations are medium for each step of the empirical cycle, with the exception of 'Hypothesis' in the PPT and 'Evaluation' in the PA, indicating that general cognitive ability reflected by CCS substantially influences performance on step level.

Relationship between subskills and steps of the empirical cycle

For further exploration of the relation between the subskills, correlations were calculated for the PPT and the aggregated PA-scores and controlled for by CCS (Table 12). The SMT did not correlate significantly with items assigned to thinking skills or to science-specific skills. However, the SMT did correlate significantly with items assigned to metacognitive skills,

Table 11. Correlations (Pearson's *r*) between scores on step level and CCS (d*f* = 121).

	CCS (general cognitive ability)	
	PPT	PA
Research question	.48*	.39*
Design experiment	.33*	.42*
Hypothesis	.28**	.47*
Measure and record	.57*	.48*
Analyze	.57*	.38
Conclusion	.46*	.42*
Evaluation	–	.27***

*Correlations are significant at $p < .001$ (2-tailed); **$p = .002$; ***$p = .003$.

Table 12. Correlations (Pearson's *r*) between subskills, controlled for CCS (d*f* = 120).

	PPT thinking	PPT science	PA thinking	PA science	PA meta	Jr. MAI
PPT thinking						
PPT science	.40*					
PA thinking	.29**	.20**				
PA science	.38*	.45*	.45*			
PA meta	.35*	.17	.44*	.28**		
Jr. MAI	−.14	−.03	.02	.03	.11	
SMT	−.04	.01	.13	.17	.21**	.65*

*Correlation is significant at $p < .001$ (2-tailed); **Correlations are significant at $p < .05$ (2-tailed).

Table 13. Correlations (Pearson's *r*) between steps of the empirical cycle, controlled for by CCS (d*f* = 120).

	1	2	3	4	5	6	7	8	9	10	11	12
1. PPT research question												
2. PPT design	.03											
3. PPT hypothesis	.37*	.03										
4. PPT measure and record	.30*	.14	.16									
5. PPT analyze	.27*	.10	.35*	.32*								
6. PPT conclusion	.14	.15	.27**	.25**	.47**							
7. PA research question	.22**	.12	.19**	.17	.18	.02						
8. PA design	.20**	.05	.22**	.39*	.31*	.17	.25*					
9. PA hypothesis	.06	.04	.21**	.13	.24*	.20**	.25*	.19**				
10. PA measure and record	.04	−.03	.26**	.41*	.26*	.30*	.24*	.41*	.26*			
11. PA analyze	.14	.14	.19**	.13	.21**	.15	.33*	.25*	.26*	.26*		
12. PA conclusion	.16	−.03	.14	.15	.23**	.09	.38*	.32*	.18**	.21**	.57*	
13. PA Evaluation	.07	.23**	.34*	.17	.27*	.06	.16	.29*	.25*	.18**	.36*	.37*

*Correlation is significant at $p < .001$ (2-tailed); **Correlations are significant at $p < .05$ (2-tailed).

indicating that thinking and science-specific items measure other skills than do the metacognitive items.

Correlations between the different steps of the empirical cycle were calculated for the PPT and the aggregated PA scores controlling for CCS (Table 13).

The positive, significant correlations between all steps of the PAs and between most corresponding steps of the PPT and PAs indicate mutual cognitive demands of the activities. In contrast, correlations between steps of the PPT are more erratic. Differences may be explained by the productive application of skills required for PAs and some items in the PPT,

in contrast to the more receptive way (students are asked to choose between alternative answers) skills are applied in most items of the PPT. This can be illustrated by the small correlation of .05 between the PPT and PAs concerning the activity of designing an experiment.

Conclusion and discussion

The present study shows that science skills can be measured reliably in grades 5 and 6. By categorizing items systematically on subskill and step level, sufficient reliabilities for the different science measures (PPT, PAs, and metacognitive self-report tests) can be obtained. The results of previous research (cf. Pine et al. 2006; Roberts and Gott 2006), showing that students' ability to perform a scientific inquiry is significantly related to their general cognitive ability, are reaffirmed in this study provided the combined Cito scores are interpreted as a reflection of general cognitive ability. Correlations between CCS and the PPT and PAs varied between .42 and .67 indicating that – despite the fact that some overlap still exists with general cognitive ability – a different construct is being measured. This implies that the tests primarily measure skills other than general cognitive ability.

Former research gave ample evidence that convergent validity between different tests for measuring science skills is difficult to establish (cf. Pine et al. 2006; Hammann et al. 2008). This lack of convergence between tests intending to measure similar science skills suggests that items within these tests do not equally appeal to underlying cognitive abilities. As demonstrated in this study, categorization of items in relation to science-specific, thinking and metacognitive skills results in more systematic test construction and thus provides evidence for convergent validity. In addition, using the steps (activities) of the empirical cycle as a blueprint ensures that within tests all aspects of scientific inquiry are incorporated (Messick 1994; Mislevy and Haertel 2006). The added value of this two-way approach is confirmed by the significant correlations between measurement instruments found in this study. This shows that lack of convergence between tests can be reduced. It should be emphasized that this applies to the relation between the PTT and PAs but also to the mutual relationship between the PAs. Although differences in difficulty level between PAs exist, the significant correlations provide evidence that inconsistencies as reported in prior studies (i.e. Pine et al. 2006) can be reduced considerably and that the problem of occasion sampling variability can be tackled by administering more than one PA. The implication is that for reliable assessment of science skills, the implementation of multiple PAs should be considered. Also, instead of using one assessment format to assess students' performance of a scientific inquiry, a greater variety of test formats may provide a clearer picture of students' abilities (Gott and Duggan 2002).

Previous studies showed that metacognitive skills have a positive influence on performing scientific inquiry (White and Frederiksen 1998). In this study, no relations were found between the Jr. MAI, measuring general metacognitive ability, the PPT and all three PAs. By contrast the SMT, constructed to obtain specific information about application of metacognitive abilities in performing science tasks, shows significant – although small – correlations with two PAs. This indicates that it is preferable to assess metacognition in performing scientific inquiry with items that are related to metacognitive activities in which students have a clear understanding of both science context and task. For young children, this may be especially essential.

The low or even lack of consistency between students' ability in performing scientific inquiry and their metacognitive self-assessment may be explained by the fact that students overestimate their own metacognitive skills. The scores on both the Jr. MAI and the SMT reveal that most students assess their own level of metacognition above the scale mean. It is therefore conceivable that many students in grades 5 and 6 are not yet able to utilize these metacognitive abilities while performing science tasks or, alternatively, simply do not master these skills even though they think they do. The latter seems most likely, given the low scores on the three items measuring specific metacognitive activities in the PAs (see Table 7). This is in line with Veenman, Van Hout-Wolters, and Afflerbach (2006) who argued that scores on questionnaires 'hardly correspond to actual behavioral measures during task performance'. This is also consistent with the science curriculum in the Netherlands in which little to no attention is being paid to the acquisition of metacognitive skills in science lessons. The implication is that – when students do not yet show possession of the metacognitive skills with which to assess their own capabilities – it may be more appropriate to use other measurement methods such as thinking aloud methods.

Assessment in primary schools is dominated by recall of procedural knowledge and the practicalities of an inquiry but typically neglects the critical evaluation of results and own performance of the tasks (Osborne and Dillon 2008; Osborne 2014; Roth 2014). By systematically including a more diverse set of items appealing to all cognitive abilities in both PPT and PAs a more valid representation of all aspects of scientific inquiry was obtained. In particular in the PAs, students obtained data by handling materials representing the practical aspects of the scientific inquiry. However, the larger part of the PAs included items in which students analyzed their own data and evaluated their own findings and performance, reflecting aspects of all three domains of activities (NRC 2012).

The last research question concerned the extent to which the measurements may provide additional diagnostic information on subskill and step level. To that end, each activity performed in a scientific inquiry was classified by determining the primary skill underlying the activity. Although all subskills were applied in the assessments in an integrated manner (van Merriënboer, Clark, and de Croock 2002), correlations between mean scores of each subskill scale – consequently reflecting the main underlying skill applied in that particular cluster of activities – between the different measures indicate that a more precise identification of students' ability in performing scientific inquiry may be allowed. The acceptable scale reliabilities on subskill level for the PPT or aggregated across PAs indicate that scores can be used to obtain diagnostic information in addition to overall test scores. To illustrate, when scores on subskill level of the PPT (consisting primarily of multiple-choice items) are compared with the PAs, the scores on subskill level in the PAs are lower. This may indicate that it is more difficult for students to report their findings and formulate their own answers than to choose between alternative answers. This is also demonstrated by the small to medium correlations on subskill level between the different assessments (PPT and PA) indicating that the assessments on subskill level may differ. Also, concerning the PAs, it seems that on average students have more difficulty in completing items in which primarily thinking and metacognitive skills underlie the activities, compared to science-specific skills.

The systematic assembly of all aspects of a scientific inquiry can also create opportunities for evaluating students' scores at a more precise level. Within tests, differences between students' performance of the different activities of the empirical cycle are manifest. For instance, results on empirical step level show that students appeared to have more difficulty

in designing an experiment than in formulating a research question. Moreover, scores for 'measure and record' were low in both the PPT and PA, as were the scores for 'analyze' and 'conclusion' in the PA. A possible explanation for these findings may be that in both the PPT and PA, the step of 'measure and record' included items in which students had to make a table and a graph. Being novices in performing a scientific inquiry, students most likely did not have the procedural knowledge to complete these items.

Nevertheless, it can be suggested that the operationalization of subskills and activities in the present study is rather indefinite. The subskill thinking, for example, still comprises a variety of mental processes such as problem-solving, making decisions, or creative thinking. And although items concerning, for example, designing an experiment were mainly categorized as science-specific based on the criterion that science-specific skills prevail in this activity, thinking and metacognitive skills are involved as well. This notion could also explain the small to medium correlations that exist between the different subskill scales within the assessments. This, together with the relatively low Cronbach α coefficients on subskill and step level, implies that estimating students' development on scale level should be made with caution. A measurement with more items specifically aimed at only thinking or other skills may improve test validity and reliability but carries the risk of becoming too detailed. Assessing all single aspects of the science skills separately may not have the same quality as assessing all aspects together in an integrated manner (Moseley et al. 2005), or in other words, the whole is more than the sum of its parts.

The constrained and scaffolded PAs seem suitable for students in primary education who are novices in performing scientific inquiry and do not yet master the required skills, although previous research shows that there is much variation between students in what they are capable of in terms of performing a scientific inquiry (Duschl, Schweingruber, and Shouse 2007). The relatively low score means of the PAs suggest that, when exposed to actual teaching and practice of science skills, the PAs have potential to measure progress of students' skill ability. Furthermore, the added value of the measures is that by structuring items by subskill and step level, the opportunity to see how students perform on a cognitive and more detailed level is provided, as opposed to merely holistically assessing performance of a scientific inquiry. Such measures are not only suitable for assessing students' mastery level of science skills, but they may provide teachers with additional diagnostic information to adapt their instructions and foster the learning process of their students.

These measures may also stimulate teachers to implement assessments in their classrooms. Teachers are mostly concerned with science activities in the curriculum, often neglecting to assess what students have learned during these activities (Harlen 1991). To some extent, this may be attributed to lack of confidence with the use of more extensive tasks than with the easy-to-administer-and-grade tests containing primarily multiple-choice items (Harlen 1991). Using a constrained PA may be less of an obstacle for teachers because of the more structured design and layout of the test. In addition, the particular format provides the opportunity to implement only parts of the PAs so that testing can be spread over more than one occasion. Moreover, the PAs can be embedded in science lessons as part of instruction material. It can be a start for familiarizing teachers with alternative assessment formats and may lead to greater confidence to implement more interesting and open inquiry tasks as students develop more skill expertise. Such tasks may eventually include aspects of scientific inquiry which are more complex and demand more of students' skill proficiency and amount of content knowledge. For instance, asking students to engage in argumentation

about different experimental designs and connect their findings to bigger ideas in science (Osborne 2014).

In primary education, the acquisition of science skills is generally measured without systematically taking into account the complexity of underlying cognitive demands that students need to simultaneously apply in relation to different activities when conducting a scientific inquiry. Categorizing items on both subskill and step level provides more opportunities for systematic test construction and improves concurrence of measurement instruments with different key content and formats such as a PPT and a PA. Furthermore, identifying and separating the various cognitive demands in assessments can help to evaluate and subsequently remedy the shortcomings of the particular skills and may also increase the emphasis in classrooms on the minds-on part of a scientific inquiry (Kind 2013). As argued by Roth (2014), assessment of skills is important because it assures that skills are taught.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by a grant from the National Platform Science & Technology [Platform Bèta Techniek] in the context of a Call for Proposals Science Skills (2015).

ORCID

Patricia M. Kruit http://orcid.org/0000-0002-3734-1904
Ron J. Oostdam http://orcid.org/0000-0003-4701-0153
Jaap A. Schuitema http://orcid.org/0000-0002-0125-5537

References

Abrahams, I., and M. J. Reiss. 2015. "The Assessment of Practical Skills." *School Science Review* 96 (357): 40–44.

Ahlbrand, W., W. Green, J. Grogg, O. Gould, and D. A. Winnett. 1993. *Science Performance Assessment Handbook*. Edwardsville, IL: Science Teachers Association.

Alonzo, A. C., and P. R. Aschbacher. 2004, April. "Value-added? Long Assessment of Students' Scientific Inquiry Skills." In *Proceedings of Assessment for Reform-based Science Teaching and Learning*. Symposium conducted at the annual meeting of the AERA. San Diego, CA.

Bartels, M., M. J. H. Rietveld, G. C. M. Van Baal, and D. I. Boomsma. 2002. "Heritability of Educational Achievement in 12-year-olds and the Overlap with Cognitive Ability." *Twin Research and Human Genetics* 5 (6): 544–553.

Baxter, G. P., R. J. Shavelson, S. R. Goldman, and J. Pine. 1992. "Evaluation of Procedure-Based Scoring for Hands-On Science Assessment." *Journal of Educational Measurement* 29 (1): 1–17.

Baxter, G. P., and R. J. Shavelson. 1994. "Science Performance Assessments: Benchmarks and Surrogates." *International Journal of Educational Research* 21 (3): 279–298.

Bloom, B. S., ed. 1956. *Taxonomy of Educational Objectives: Handbook 1, Cognitive Domain*. New York: David McKay.

Clauser, B. E. 2000. "Recurrent Issues and Recent Advances in Scoring Performance Assessments." *Applied Psychological Measurement* 24 (4): 310–324.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Davey, T., S. Ferrara, R. Shavelson, P. Holland, N. Webb, and L. Wise. 2015. *Psychometric Considerations for the Next Generation of Performance Assessment*. Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service.
Donovan, M. S., J. D. Bransford, and J. W. Pellegrino. 1999. *How People Learn: Bridging Research and Practice*. Washington, DC: National Academies Press.
Duschl, R. A., H. A. Schweingruber, and A. W. Shouse. 2007. *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: National Academies Press.
Eberbach, C., and K. Crowley. 2009. "From Every Day to Scientific Observation: How Children Learn to Observe the Biologist's World." *Review of Educational Research* 79 (1): 39–68.
Ennis, R. H. 1993. "Critical Thinking Assessment." *Theory Into Practice* 32 (3): 179–186.
Flavell, J. H., P. H. Miller, and S. A. Miller. 1993. *Cognitive Development*. Upper Saddle River, NJ: Prentice-Hall.
Fraser, B. J. 1979. *Test of Enquiry Skills [and] Handbook*. Hawthorn: Australian Council for Educational Research.
Fraser, B. J. 1980. "Development and Validation of a Test of Enquiry Skills." *Journal of Research in Science Teaching* 17 (1): 7–16.
Gobert, J. D., and K. R. Koedinger. 2011, September. "Using Model-tracing to Conduct Performance Assessment of Students' Inquiry Skills Within a Microworld." Paper presented at the Society for Research on Educational Effectiveness, Washington, DC.
Goodson, L. A. 2000. "Teaching and Learning, Strategies for Complex Thinking Skills." *In Annual Proceedings of Selected Research and Development Papers* 1 (2): 164–172.
Gott, R., and S. Duggan. 1995. *Investigative Work in the Science Curriculum*. Buckingham: Open University Press.
Gott, R., and S. Duggan. 2002. "Problems with the Assessment of Performance in Practical Science: Which way now?" *Cambridge journal of education* 32 (2): 183–201.
Gott, R., and P. Murphy. 1987. *Assessing Investigation at Ages 13 and 15: Assessment of Performance Unit Science Report for Teachers: 9*. London: Department of Education and Science.
Hammann, M., T. T. H. Phan, M. Ehmer, and T. Grimm. 2008. "Assessing Pupils' Skills in Experimentation." *Journal of Biological Education* 42 (2): 66–72.
Harlen, W. 1986. *Science at age 11: Assessment of Performance Unit Science Report for Teachers: 1*. London: Department of Education and Science.
Harlen, W. 1991. "Pupil Assessment in Science at the Primary Level." *Studies in Educational Evaluation* 17 (2–3): 323–340.
Harlen, W., and A. Qualter. 2009. *The Teaching of Science in Primary Schools*. Abingdon: Routledge.
Harmon, M., T. A. Smith, M. O. Martin, D. L. Kelly, A. E. Beaton, I. Mullis, E. J. Gonzalez, and G. Orpwood. 1997. *Performance Assessment: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.
Hofstein, A., and V. N. Lunetta. 2004. "The laboratory in Science Education: Foundations for the Twenty-First Century." *Science Education* 88 (1): 28–54.
Janssen, J. N., D. Verhelst, R. J. H. Engelen, and F. Scheltens. 2010. *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8* [Scientific Justification of Tests Arithmetic-Math for Grades 1 to 6]. Arnhem: Cito.
Jones, L. R., G. Wheeler, and V. A. S. Centurino. 2013. "TIMSS 2011 Science Framework." In *TIMSS 2015 Assessment Frameworks*, edited by I. V. S. Mullis and M. O. Martin, 49–90. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
de Jong, P. F., and E. A. Das-Smaal. 1995. "Attention and Intelligence: The Validity of the Star Counting Test." *Journal of Educational Psychology* 87 (1): 80–92.
Kane, M., T. Crooks, and A. Cohen. 1999. "Validating Measures of Performance." *Educational Measurement: Issues and Practice* 18 (2): 5–17.
Kind, P. M. 1999. "Performance Assessment in Science - What are We Measuring?" *Studies in Educational Evaluation* 25 (3): 179–194.
Kind, P. M. 2013. "Establishing Assessment Scales Using a Novel Disciplinary Rationale for Scientific Reasoning." *Journal of Research in Science Teaching* 50 (5): 530–560.
Kuhn, D. 1989. "Children and Adults as Intuitive Scientists." *Psychological Review* 96 (4): 674–689.

Kuhn, D. 1999. "A Developmental Model of Critical Thinking." *Educational Researcher* 28 (2): 16–46.
Kuhn, D., and D. Dean Jr. 2004. "Metacognition: A Bridge Between Cognitive Psychology and Educational Practice." *Theory Into Practice* 43 (4): 268–273.
Lawrenz, F., D. Huffman, and W. Welch. 2001. "The Science Achievement of Various Subgroups on Alternative Assessment Formats." *Science Education* 85 (3): 279–290.
Lawson, A. E. 1989. "Research on Advanced Reasoning, Concept Acquisition and a Theory of Science Instruction." In *Adolescent Development and School Science*, edited by P. Adey, 11–36. London: Falmer Press.
Lederman, N., and J. Lederman. 2014. "Research on Teaching and Learning of Nature of Science." In *Handbook of Research on Science Education, 2*, edited by S. K. Abell and N. G. Lederman, 600–620. Abingdon: Routledge.
Lewis, A., and D. Smith. 1993. "Defining Higher Order Thinking." *Theory Into Practice* 32 (3): 131–137.
Martin, M. O., I. V. Mullis, A. E. Beaton, E. J. Gonzalez, T. A. Smith, and D. L. Kelly. 1997. *Science Achievement in the Primary School Years IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
van Merriënboer, J. J. G., R. E. Clark, and M. B. M. de Croock. 2002. "Blueprints for Complex Learning: The 4C/ID-model." *Educational Technology Research and Development* 50 (2): 39–61.
Messick, S. 1994. "The Interplay of Evidence and Consequences in the Validation of Performance Assessments." *Educational Researcher* 23 (2): 13–23.
Millar, R., and R. Driver. 1987. "Beyond Processes." *Studies in Science Education* 14 (1): 33–62.
Mislevy, R. J., and G. D. Haertel. 2006. "Implications of Evidence-centered Design for Educational Testing." *Educational Measurement: Issues and Practice* 25 (4): 6–20.
Moseley, D., V. Baumfield, J. Elliott, M. Gregson, S. Higgins, J. Miller, and D. P. Newton. 2005. *Frameworks for Thinking: A Handbook for Teaching and Learning*. Cambridge: Cambridge University Press.
Newmann, F. M. 1990. "Higher Order Thinking in Teaching Social Studies: A Rationale for the Assessment of Classroom Thoughtfulness." *Journal of Curriculum Studies* 22 (1): 41–56.
te Nijenhuis, J., E. Tolboom, W. Resing, and N. Bleichrodt. 2004. "Does Cultural Background Influence the Intellectual Performance of Children from Immigrant Groups?: Validity of the RAKIT Intelligence Test for Immigrant Children." *European Journal of Psychological Assessment* 20 (1): 10–26.
National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.
OECD. 2017. *PISA for Development Assessment and Analytical Framework*. Paris: OECD.
Osborne, J. 2014. "Teaching Scientific Practices: Meeting the Challenge of Change." *Journal of Science Teacher Education* 25 (2): 177–196.
Osborne, J. 2015. "Practical Work in Science: Misunderstood and Badly Used?" *School Science Review* 96 (357): 16–24.
Osborne, J., and J. Dillon. 2008. *Science Education in Europe: Critical Reflections*. vol. 13. London: The Nuffield Foundation.
Pine, J., P. Aschbacher, E. Roth, M. Jones, C. McPhee, C. Martin, S. Phelps, T. Kyle, and B. Foley. 2006. "Fifth Graders' Science Inquiry Abilities: A Comparative Study of Students in Hands-on and Textbook Curricula." *Journal of Research in Science Teaching* 43 (5): 467–484.
Pintrich, P. R. 2002. "The Role of Metacognitive Knowledge in Learning, Teaching, and Assessing." *Theory Into Practice* 41 (4): 219–225.
Roberts, R., and R. Gott. 2006. "Assessment of Performance in Practical Science and Pupil Attributes." *Assessment in Education: Principles, Policy & Practice* 13 (1): 45–67.
Roth, K. J. 2014. "Elementary Science Teaching." In *Handbook of Research on Science Education*. vol. 2, edited by N. G. Ledermann and S. K. Abell, 361–393. New York: Routledge.
Ruiz-Primo, M. A., G. P. Baxter, and R. J. Shavelson. 1993. "On the Stability of Performance Assessments." *Journal of Educational Measurement* 30 (1): 41–53.
Schilling, M., L. Hargreaves, W. Harlen, and T. Russell. 1990. *Assessing Science in the Primary Classroom: Written Tasks*. London: Paul Chapman Publishing.
Schraw, G., K. J. Crippen, and K. Hartley. 2006. "Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning." *Research in Science Education* 36 (1–2): 111–139.

Schraw, G., and D. Moshman. 1995. "Metacognitive Theories." *Educational Psychology Review* 7 (4): 351–371.
Shavelson, R. J., G. P. Baxter, and J. Pine. 1991. "Performance Assessment in Science." *Applied Measurement in Education* 4 (4): 347–362.
Shavelson, R. J., N. B. Carey, and N. M. Webb. 1990. "Indicators of Science Achievement: Options for a Powerful Policy Instrument." *Phi Delta Kappan* 71 (9): 692–697.
Shavelson, R. J., G. Solano-Flores, and M. A. Ruiz-Primo. 1998. "Toward a Science Performance Assessment Technology." *Evaluation and Program Planning* 21 (2): 171–184.
Solano-Flores, G., J. Javanovic, R. J. Shavelson, and M. Bachman. 1999. "On the Development and Evaluation of a Shell for Generating Science Performance Assessments." *International Journal of Science Education* 21 (3): 293–315.
Song, J., and P. J. Black. 1992. "The Effects of Concept Requirements and Task Contexts on Pupils' Performance in Control of Variables." *International Journal of Science Education* 14 (1): 83–93.
Sperling, R. A., B. C. Howard, L. A. Miller, and C. Murphy. 2002. "Measures of Children's Knowledge and Regulation of Cognition." *Contemporary Educational Psychology* 27 (1): 51–79.
Steiger, J. H. 1980. "Tests for Comparing Elements of a Correlation Matrix." *Psychological Bulletin* 87: 245–251.
Sternberg, R. J. 1985. *Beyond IQ: A Triarchic Theory of Human Intelligence*. New York: Cambridge University Press.
Tamir, P. 1988. "Science Practical Process Skills of Ninth Grade Students in Israel." *Research in Science & Technological Education* 6 (2): 117–131.
Veenman, M. V. J., B. H. Van Hout-Wolters, and P. Afflerbach. 2006. "Metacognition and Learning: Conceptual and Methodological Considerations." *Metacognition and Learning* 1 (1): 3–14.
Veenman, M. V. J., P. Wilhelm, and J. J. Beishuizen. 2004. "The Relation Between Intellectual and Metacognitive Skills from a Developmental Perspective." *Learning and Instruction* 14 (1): 89–109.
Weekers, A., I. Groenen, F. G. M. Kleintjes, and H. Feenstra. 2011. *Wetenschappelijke verantwoording papieren toetsen begrijpend lezen voor groep 7 en 8* [Scientific Justification Paper-and-Pencil Tests Reading Comprehension for Grades 5 and 6]. Arnhem: Cito.
White, B. Y., and J. R. Frederiksen. 1998. "Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students." *Cognition and Instruction* 16 (1): 3–118.
White, B. Y., and J. R. Frederiksen. 2000. "Metacognitive Facilitation: An Approach to Making Scientific Inquiry Accessible to All." In *Inquiring into Inquiry Learning and Teaching in Science*, edited by J. Minstrell and E. van Zee, 331–370. Washington, DC: American Association for the Advancement of Science.
Zohar, A., and S. Barzilai. 2013. "A Review of Research on Metacognition in Science Education: Current and Future Directions." *Studies in Science Education* 49 (2): 121–169.
Zohar, A., and Y. J. Dori. 2003. "Higher Order Thinking Skills and Low-Achieving Students: Are They Mutually Exclusive?" *Journal of the Learning Sciences* 12 (2): 145–181.

Appendix 1. Example items of paper-and-pencil test

Mustafa puts four leaves in a container together with a caterpillar. He checks the leaves every day. What can he tell from what is happening with the leaves every day?

Monday	Tuesday	Wednesday	Thursday	Friday

A. The caterpillar only eats every other day.
B. The caterpillar eats every single day.
C. The caterpillar will eat a piece of a leave the next Monday.
D. The caterpillar will stop eating on Friday when he is going to pupate.

Figure A1. Item in which students make inferences informed by evidence and reason, assigned to thinking. Source: SOLpass.org.

In the figure you can see a thermometer. The thermometer is hanging in a room. What is the temperature in this room?

A. 19^0 C
B. 20^0 C
C. 21^0 C
D. 22^0 C

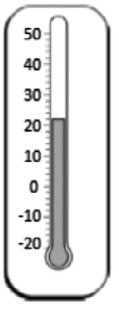
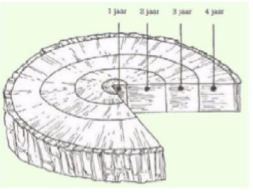
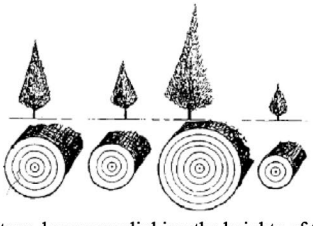


Figure A2. Example item assigned to science-specific: observe/measure correctly. Source: Fraser (1979).

When we cut across the trunk of a tree we see growth rings:



The trees below were planted at different times in the same wood. The drawings underneath show the growth rings seen when the trees were cut down.



What pattern do you see linking the heights of the trees and the rings in the trunk?
Write your answer below:

Figure A3. Open-ended item in which students identify patterns, assigned to thinking. Source: Harlen (1986).

A group of children wanted to find out how fast a plant grows. For 6 days in a row the children kept a record of how much the plant grew. They saw that the plant grew 1,5 cm on each day. The children put the results in a table. What does the table look like? Make a table and note the results.

Figure A4. Example open-ended item in which students record data, assigned to science-specific.

Appendix 2. Example of scoring model for making a graph

For both labeling the axes and drawing the line, points are awarded and subsequently added. Fill out this score.

0	No numbers at axes/numbers count backwards on one of both axes	
1	-numbers are (linearly) and evenly spread on axes -numbers start with 0* (if not mentioned, but space is made) -large part of space used (more than half the space available)	2 out of 3
2	Numbers are (linearly) and evenly spread on axes; large part of space used (more than half the space available); numbers start with 0* (if not mentioned, but space is made); all measurements fall into space available	

0	A bar chart is made/there are no data points
1	A line is drawn, but the data points are not correct / data points are correct, but no line is drawn
2	All data points are correct; a line is drawn through the data points / there is a line of best-fit

Goal:
The student is capable of drawing a graph. The numbers of the axes are correct (starting with zero of space left open in first cell) and the data points are noted correctly. A line has been drawn through the data points. The graph uses most of the space available.
Notes:

- No points are deducted when helping lines are drawn.
- If a random line is drawn without taking into account the data points: no points are awarded for line.
- If numbers are not on the lines of axes, but instead between lines, no points are deducted.

*numbering must be sequential between zero and the next number. For instance: (-)-5-10-15 or (-)-2-4-6. And not: (-)-50-55-60, unless specifically a cell is left open.

Appendix 3. Example of scoring rubric of formulating a research question in performance assessment Skateboard

Item 1 Research question	Can you think of a research question you want to find an answer to? Write down your question:	
0 Leaves space empty/formulates a question not relevant, understandable or possible to investigate/ just refers to illustration	*Who is right?* *Who rolls further?* *How can they go faster?*	Goal: The student is able to formulate a research question relating to the goal of the investigation. Goal of the investigation is to find the relation between the distance on the ruler and the distance the marble covers at the end of the ruler. The research question is relevant if it leads to finding this relation.
1 Formulates a researchable question which can be answered with results of this experiment, but has no connection to relationship between distance on ruler and distance marble rolling. (Question is on itself understandable and relates to the context of skating (or marbles))	*Can the marble roll a distance of 15 cm?* *How far can the marble push the paper wedge?* *Do you go faster when you start higher up the hill?*	Notes: If a relation is mentioned, but speed is included, only 1 point is assigned In case formulation leads to answering yes/no: no points are subtracted
2 Formulates a researchable question which can be answered with results of this experiment and (explicitly) identifies the relationship between distance on ruler and distance of marble rolling	*Does the marble roll further when the marble starts higher up the hill then when the marble starts at a lower point?* *Do you go further when you start higher up the hill?*	