Performance assessment as a diagnostic tool for science teachers

Kruit, P.; Oostdam, R.; van den Berg, E.; Schuitema, J.

DOI
10.1007/s11165-018-9724-9

Publication date
2020

Document Version
Final published version

Published in
Research in Science Education

License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Kruit, P., Oostdam, R., van den Berg, E., & Schuitema, J. (2020). Performance assessment as a diagnostic tool for science teachers. *Research in Science Education*, *50*(3), 1093–1117. https://doi.org/10.1007/s11165-018-9724-9

Res Sci Educ (2020) 50:1093–1117
https://doi.org/10.1007/s11165-018-9724-9

Performance Assessment as a Diagnostic Tool for Science Teachers

Patricia Kruit[1] **· Ron Oostdam**[1,2] **· Ed van den Berg**[1] **·**
Jaap Schuitema[2]

Published online: 18 April 2018

Abstract Information on students' development of science skills is essential for teachers to evaluate and improve their own education, as well as to provide adequate support and feedback to the learning process of individual students. The present study explores and discusses the use of performance assessments as a diagnostic tool for formative assessment to inform teachers and guide instruction of science skills in primary education. Three performance assessments were administered to more than 400 students in grades 5 and 6 of primary education. Students performed small experiments using real materials while following the different steps of the empirical cycle. The mutual relationship between the three performance assessments is examined to provide evidence for the value of performance assessments as useful tools for formative evaluation. Differences in response patterns are discussed, and the diagnostic value of performance assessments is illustrated with examples of individual student performances. Findings show that the performance assessments were difficult for grades 5 and 6 students but that much individual variation exists regarding the different steps of the

✉ Patricia Kruit
p.m.kruit@hva.nl

Ron Oostdam
r.j.oostdam@hva.nl

Ed van den Berg
e.berg@vu.nl

Jaap Schuitema
j.a.Schuitema@uva.nl

[1] Centre for Applied Research in Education, Amsterdam University of Applied Sciences, Wibautstraat 2-4, 1091 GM Amsterdam, Netherlands

[2] Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, Netherlands

empirical cycle. Evaluation of scores as well as a more substantive analysis of students' responses provided insight into typical errors that students make. It is concluded that performance assessments can be used as a diagnostic tool for monitoring students' skill performance as well as to support teachers in evaluating and improving their science lessons.

Keywords Performance assessment · Formative assessment · Science education · Science skills · Individual differences · Primary school

Introduction

With the increased attention towards the implementation of inquiry activities within primary science classrooms, educators have emphasized the importance of including assessments in science classroom practice (Duschl et al. 2007; National Research Council (NRC) 2012). Information relating to students' development of science skills is essential for teachers to evaluate and improve their own education, as well as to provide adequate support to the learning process of individual students (Germann and Aram 1996; Roth 2014). In addition, including assessments of science skills is important because it ensures that these skills are taught (Harlen et al. 2012; Vogler 2002). Performance assessments have been considered to be an alternative to paper-and-pencil tests for the assessment of science skills (Harmon et al. 1997; NRC 2012; Shavelson et al. 1991). In a performance assessment, students perform small investigations by interacting with real materials. Students' performance is evaluated on the basis of the actions they take and subsequently report on regarding their investigations. The present study explores and discusses the use of performance assessments as a diagnostic tool for formative assessment to inform teachers and guide instruction of science skills in primary education.

Skills and Scientific Inquiry

Science skills—also referred to as inquiry skills, process skills, or investigation skills (Harlen and Qualter 2009)—usually indicate a wide variety of activities related to planning and conducting investigations and interpreting results (Alonzo and Aschbacher 2004; Gott and Duggan 1995; Harlen and Qualter 2009). In current literature, it is generally acknowledged that scientific inquiry activities in science classrooms should be based on the actual work of scientists (Lederman and Lederman 2014; Pellegrino 2014). Within the framework of K-12 science education, the NRC (2012) argues that students should learn about what actual scientists do when designing and carrying out inquiries. One of the aims is to understand how knowledge about issues such as health and environment is obtained and validated.

Authentic research involves three domains of activities in which scientists go back and forth: investigating, developing explanations and solutions, and evaluating data as evidence for the proposed theories and models (NRC 2012). This implies that in performing a scientific inquiry, a wide variety of cognitive abilities are invoked. For example, different abilities are employed when handling a microscope than when identifying patterns in data. Consequently, the consideration of different skill categories underlying the inquiry activities has been generally acknowledged (cf. Schraw et al. 2006; Zohar and Dori 2003). In particular, three categories which underlie performing a scientific inquiry may be involved: *science-specific* skills, *thinking* skills, and *metacognitive* skills (see also Kruit et al. 2018a).

Science-specific skills can be classified as lower-order thinking (Newmann 1990) which is defined by knowledge recall, routine employment, and simple application of rules (Goodson 2000). These skills include practical skills such as taking measurements and using a microscope (Abrahams and Reiss 2015) but relate to cognitive processes as well. That is, in performing a scientific inquiry, students must recall the facts and rules which are specific for the science domain and then apply this knowledge in the appropriate manner (Gott and Murphy 1987; OECD 2017). For example, converting data into tables and graphs can be regarded as a science-specific skill.

Thinking skills include the higher-order skills, also frequently referred to as critical thinking (Moseley et al. 2005). Thinking skills involve manipulating complex information which consists of more than one element and has a high level of abstraction (Bloom 1956; Flavell et al. 1993). In a scientific inquiry, thinking skills are applied to make sense of the data and connect the observations to scientific theories (Osborne 2015). These include, for example, formulating hypotheses, interpreting, analyzing, and evaluating data, drawing a conclusion, and classifying and inferring information (Moseley et al. 2005; Newmann 1990; Pintrich 2002; Zohar and Dori 2003). Also, metacognitive skills such as planning, monitoring, and evaluating task performance are considered key in promoting the quality of scientific inquiry (Schraw et al. 2006). For instance, evaluating their plan while performing the inquiry helps students to rethink what they are doing and adjust the plan if necessary (Michaels et al. 2007). As argued by Kuhn (1997), the essence of scientific thinking is coordinating theory and evidence which specifically demands metacognitive skills.

For science education, the influence of content knowledge on skill development is generally considered to be of paramount importance (Duschl et al. 2007). Content knowledge is generally referred to as a conceptual understanding of facts, concepts, theories and principles (OECD 2017). Many studies have shown that the level of content knowledge has substantial impact on skill development (Eberbach and Crowley 2009; Kuhn et al. 1992). Particularly, prior content knowledge influences the quality of students' inquiry performance when they generate hypotheses, make observations, evaluate evidence, and draw conclusions (Duschl et al. 2007; Millar and Driver 1987).

Performance Assessment

As an alternative to standardized paper-and-pencil tests, performance assessments (PAs) are considered to be valid instruments when assessing students' science skills (Shavelson et al. 1991; NRC 2012). In PAs, students perform small experiments by interacting with real materials. The small experiments in PAs are typically organized according to the different steps of the empirical cycle which include: (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. As is generally acknowledged, scientists do not follow these steps linearly during actual research (NRC 2012). However, the steps provide a structure that is recognizable for students which is particularly important for students in primary education who have little experience with scientific inquiry (Donovan et al. 1999). Therefore, the different steps provide a suitable framework for systematic (formative) evaluation of students' science skills.

A PA generally consists of three components: a task, a response demand, and a scoring system (Shavelson et al. 1998). A PA can be considered to be a test on a particular topic which contains a set of items. The main characteristics of what defines a PA is that it is a complex task set in a context reflecting real-life experiences and in which different skills and knowledge are interconnected (Davey et al. 2015). The response demand may be verbal which requires observation measures in order to properly score it. It may also be written, for example, by means of a worksheet or a notebook. For measurement of skills involved in a scientific inquiry, a scientific report may be considered a valid reflection of authentic research since scientists use reports to communicate their findings. The nature of a scoring system depends on the type of task used. For example, for tasks in which students use worksheets to note results and write down answers to questions, a scoring rubric may be used to rate the students' responses.

PAs have been implemented for large-scale testing such as in The Third International Mathematics and Science Study of TIMSS in 1995, and in the National Assessment of Educational Progress (NAEB) of 2009 (Britton and Schneider 2014). Research up until now has been concerned with the limitations and advantages of implementing PAs for summative assessment. Previous studies showed low correlations between different PAs designed to measure the same science skills (Gott and Duggan 2002; Pine et al. 2006). The main problem of using PAs for testing is that students perform differently on similar PAs from one occasion to the other (Ruiz-Primo et al. 1993; Solano-Flores et al. 1999). Various reasons for this occasion sampling variability have been put forward. For instance, the PAs generally differ in the context (the topic which reflects a real-life phenomenon) in which they are set. Students actual task performance depends on the knowledge they have about the topic of the task as well as how well they are able to apply that knowledge (Klassen 2006). This implies that, although the PAs call upon similar science skills, students perform differently for each PA due to the knowledge the student has of the particular topic of the PAs (Gott and Duggan 2002; Shavelson et al. 1991). Accordingly, Ruiz-Primo et al. (1993) suggested that a reliable measurement of science skills may only be obtained by administering a substantial number of PAs.

Scoring of students' responses is generally considered a substantial challenge when implementing PAs. The scoring can either be based on directly observing the performance itself or on students' written responses which are evaluated after the event (Clauser 2000). In science education, it is generally acknowledged that observing is not practical for large-scale and classroom testing (e.g., Klassen 2006). Therefore, research has been concerned with evaluating students' written answers. An important advantage of written answers is that it lends the possibility of analyzing and scoring responses after the event has taken place (Schilling et al. 1990). Because PAs elicit skills similar to those that scientists apply when they perform a scientific inquiry and subsequently report the results, scoring the answers after the event is assumed to provide a valid indication of students' potential performance in real-life inquiry (Davey et al. 2015; Harmon et al. 1997; Kind 1999).

However, it can be argued that a response format requiring extensive written answers demands a certain level of writing abilities. Harlen et al. (2012) argued that the response format may influence students' scores. The implication is that it may be more difficult to determine what exactly is being measured (Klassen 2006; Stecher et al. 2000). On the other hand, in a study by Baxter and Shavelson (1994) addressing the exchangeability of observational and written scoring, results indicated that scoring written responses provides a satisfactory alternative to observation. It is nevertheless important that when developing PAs, attention should be paid to the "verbal demands of tasks" (Stecher et al. 2000, p. 154).

It has also been noted that scoring criteria and rubrics may be difficult to interpret by raters and therefore require extensive training to ensure that rating between raters and occasions is consistent (Davey et al. 2015). According to Clauser (2000), reliable rating is influenced by a number of different conditions such as the extent of detail of the scoring criteria and the level of training of raters. Rating may be improved when scoring rules are described in detail with examples of possible student answers. On the other hand, the scoring criteria may get too specific which limits generalizability across tasks (Messick 1994). Also, raters may be influenced by students' handwriting and turns of phrase or assess students' responses differently on different occasions. Although score variation caused by rater effects can be reduced considerably by thorough training (e.g., Ruiz-Primo et al. 1993), the training and rating procedure is cost and labor-intensive (Davey 2015; Klassen 2006).

All of this said, it is still the case that, in contrast with paper-and-pencil tests, PAs have the important advantage of measuring skills more comprehensively because students actually apply these skills to a real-life scientific inquiry (Davey et al. 2015; Ennis 1993). The issues of reliability previously discussed have less of an impact when used for formative assessment in the classroom (Harlen 1999). In addition, as argued above, a major advantage of PAs is that students perform small experiments in which they systematically follow the various steps within the empirical cycle. This provides an opportunity to separate the various steps when determining students' performance levels. Furthermore, although a reliable rating of written responses is complicated due to the open format (Davey et al. 2015), teachers who score PA items find it of considerable value to focus on particular aspects of a students' written answers rather than merely regarding them as right or wrong (Harlen 1991). Aschbacher and Alonzo (2006) argued that notebooks can reveal students' thought processes which teachers can use to guide instruction. As a result, this use of PAs can provide ample opportunity to collect information on students' performance for not just summative but also for formative evaluation.

Formative Assessment

The primary purpose of formative assessment is to observe progress made and to collect information to guide subsequent teaching and learning (Harlen et al. 2012). Formative assessment elicits evidence of student performance and thus provides the teacher with information which can be used to modify teaching and classroom activities (Black et al. 2004; Wiliam and Thompson 2007). In a science classroom, the assessments may be spontaneously incorporated in the lesson by asking questions and starting discussions, or they may take the form of planned activities which are part of the curriculum (Loxley et al. 2013). Information on science skill acquisition can be gathered by observing students while prompting them to explore their ideas and reasoning. Also, students can be asked to communicate their thinking by using drawings and writing (Harlen 1999).

When implementing PAs in the science classroom, students are provided with opportunities to use skills and are encouraged to think critically about their performance, which promotes students' learning. In particular, by structuring PAs to include the steps of the empirical cycle, more detailed information can be gathered on all aspects of a scientific inquiry. Teachers can provide adequate feedback on the students' work, engage students in metacognitive discussion about the procedures applied in the PA and give them examples of a well-performed inquiry. Finally, teachers can provide students with the techniques and the language needed to perform a scientific inquiry (Davey et al. 2015; Harlen 1999).

The Present Study

The NRC (2014) has argued that assessments other than those currently used should be implemented in classrooms to reflect the contemporary vision of science education. An important aspect of the assessments is that they should be "... specific enough to be useful for helping teachers understand the range of student responses and provide tools for helping teachers decide on next steps in instruction" (NRC 2014, p. 3). In general, in daily classroom practice, teachers will spend most of their time and effort on summative assessments rather than formatively assessing their students' progress (Black et al. 2004; Black and Wiliam 2003). Even if they do implement formative assessments, this will generally not be aimed towards improving their teaching or the learning process of the students but mainly on "deciding the level of achievement" (Harlen 1999, p. 137).

The aim of the present study is to explore and discuss in what way PAs can be used to provide teachers with diagnostic information to rate group or individual student performance. When developing PAs for summative assessment, their structure should also provide opportunities to use them for formative assessment in the classroom. As a result, a more fine-grained picture of students' acquired skills may be obtained and used by the teacher to gain information about students' learning (Pellegrino 2012). Therefore, in this study, the utility of the more specific information that may be obtained by structuring the PAs according to the different steps of the empirical cycle is discussed. Furthermore, the way PAs are structured and scored creates opportunities to monitor students' learning progress. This diagnostic information is not only important for teachers to improve their teaching but also to provide (individual) students with adequate feedback. As a result, the present study may add to the understanding of using performance assessments as a tool for formatively assessing students in science classrooms. The main research question within this present study is as follows: in what way can a PA be used as a diagnostic tool to evaluate students' progress and to guide instruction in science classroom practice? To answer this question, we discuss design principles of PAs and the corresponding scoring rubrics based upon the steps of the empirical cycle. Furthermore, we analyze and discuss the different levels of student performance on the three PAs: the mean scores, students' response patterns and illustrative examples of and trends in students' responses. Finally, we examine the relationship between the PAs per step to search for evidence for the usefulness of PAs for formative evaluation.

Method

Sample

The responses on three PAs of 403 primary students (aged 10–13, 50.9% girls) were used. Of these students, 51.6% were in grade 5 and 48.4% in grade 6. The students were from 12 primary schools located in urban areas of the Netherlands. Schools were willing to participate on the basis of several pragmatic factors including permission of school authorities, interest of teachers, willingness to do something with science, and available time.

The students in the present study were participants in an effect study with an experimental, a control, and a baseline condition (Kruit et al. 2018b). Students in the experimental condition received an intervention lasting 8 weeks in which regular explicit instruction in the use of science skills was provided during regular science class. Explicit instruction involved the teacher clarifying the rationale behind these science skills followed by examples and classroom discussions about how to apply the skills. Then, students performed a scientific inquiry in which they received guiding questions and specific feedback. In the lessons within the control condition, all aspects of explicit instruction were absent. Instead, skills were merely encountered and practiced while performing scientific investigations. Students in the baseline condition followed their regular science curriculum, meaning that they did not receive formal and structured instruction on science skills during the intervention period. The PAs were implemented as summative measurement instruments.

Description and Development of the Performance Assessments

Three tasks were developed with topics suitable for grades 5 and 6 students: *Skateboard*, *Bungee Jump*, and *Hot chocolate* (Kruit et al. 2018a). Skateboard was based on the PA "Rolling Down Hill" (Ahlbrand et al. 1993). Bungee jump and Hot chocolate were based on tasks used in TIMSS (Martin et al. 1997).

All three PAs appertain to comparative investigations in which students explore the relationship between two variables (Shavelson et al. 1998). In Skateboard, students roll a marble (the "skateboard") down a ruler (the "hill") to investigate the relationship between the distance that the marble covers on the ruler (slope) and the distance the marble covers at the end of the ruler while pushing a paper wedge forward. Similarly, students investigate in Bungee jump the changing length of a rubber band as additional weights (metal rings) are added, and in Hot chocolate, students examine the relationship between the amount of hot water and its rate of cooling.

Each PA was constructed according to the same template following the various activities (steps) of the empirical cycle (Table 1). In each PA, the topic was introduced by a description of the context of the experiment. The context for Skateboard comprised a cartoon in which skateboarding children were wondering who would roll farther at the bottom of the hill (see Appendix 1 for the entire task). A cartoon for Bungee jump represented children of different weights bungee jumping off a bridge and for Hot chocolate a scenario was described of a cup of tea or hot chocolate that was still too hot to drink. The students' first task was to formulate a research question in line with the topic presented. In the following task, students were provided with a pre-formulated research question. Based on this research question, students were asked to design an experiment. This task feature ensured that the subsequent investigation designed by students was not contaminated by a flawed research question or that the designs would not become too divergent to be properly compared. The subsequent items followed the remaining steps of the empirical cycle and referred to the pre-formulated research question (see also Table 1).

University lecturers in the field of biology and physics education assessed the items regarding clarity of formulation, activity to be measured and suitability for young students. Minor adjustments were then made to the items. To provide students with additional scaffolding, students were asked to formulate a hypothesis *after* designing the experiment whereas it is typically the other way round.

Prior to implementing the PAs as summative assessments in the effect study (Kruit et al. 2018b), 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.

Table 1 Blueprint of items of the performance assessments

Item	Description of activities	Step empirical cycle	Score
1	Formulate research question	Research question	0–1–2
2	Design experiment	Design	
	Description of experimental set-up		0–1–2–3
	Description of how results will be noted		0–1–2
3	Formulate hypothesis	Hypothesis	0–1–2
4	Note results in a table students make themselves	Measure and record	0–1–2–3
5	Make a graph	Measure and record	
	Axes		0–1–2
	Line graph		0–1–2
6	Interpret results by relating two variables	Analyze	0–1–2
7	Extrapolate results	Analyze	0–1–2
8	Draw conclusion about relationship	Conclusion	0–1–2
9	Formulate support for conclusion	Conclusion	0–1–2
10	Relate hypothesis to conclusion	Conclusion	0–1–2
11	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	0–1–2
12	Give suggestions to extend the experiment	Evaluate	0–1–2
13	Draw conclusion related to the context	Conclusion	0–1–2
14	Formulate learning gains about inquiry	Evaluate	0–1–2
	Maximum score		34

Note. Maximum score per step: research question, 2; design, 5; hypothesis, 2; measure and record, 7; analyze: 4; conclusion, 8; evaluate, 6

The Scoring Rubric

To obtain a fine-grained picture of the students' ability to perform a scientific inquiry, a rubric was developed for scoring all 14 items of the PAs. Since generally speaking, existing rubrics are specifically designed to match a particular task, a new scoring rubric for the PAs was developed (see for example rubric Appendix 2).

First, the activities were operationalized by specifying what a student's response should entail. The elements of the specifications were derived from resources in which goals and learning progressions for science skills are described (e.g., Next Generation Science Standards (NGSS) 2013; https://pals.sri.com). For example, for designing an experiment, the following goal was formulated for the scoring rubric of the PAs used in this study:

> The goal is that students learn to construct a design consisting of several elements. The design is aimed at finding a relationship between two variables and is described in a structured manner, which entails that the steps are at least in chronological order. There is sufficient information to replicate the experiment.

Next, the criteria for different levels of proficiency were formulated as detailed descriptions of the elements required to appear within students' answers. University lecturers in the field of biology and physics education were employed as content experts to assess the criteria for awarding points to the different levels of proficiency of possible answers. An answer containing more elements is considered to demonstrate a higher level of ability. Depending on the number of elements, more points are awarded (see the score column in Table 1). Primary school teachers then reviewed the criteria to determine its feasibility for grades 5 and 6 students.

Students' responses obtained from the pilots were used to evaluate and adjust the scoring criteria. Characteristic responses were added as examples to illustrate the different levels of proficiency. The criteria descriptions and examples of all three PAs were equivalent and differed only in their context-specific elements (see Appendix 2 for the scoring rubric for formulating a research question).

Within each PA items were assigned to one of the steps of the empirical cycle (Table 1). For some steps, more than one item was assigned. For example, the step "measure and record data" included two items while four items were assigned to the step "formulate a conclusion". Scores for each step were obtained by summing up the scores of the assigned items. As a result, the 14 items were reduced to separate scores for each of the seven steps of the empirical cycle.

Scoring Procedure

All students' hand-written answers to the PAs were scored after having been transcribed into typed text and grouped by item. Raters were trained to interpret the criteria as it was intended and to award points to students' responses in a consistent manner. During training sessions, the scores of each item were compared separately and interpretations of the criteria and students' responses were discussed. This enabled the fine-tuning of the criteria. After establishing satisfactory interrater reliability for the total score of a random sample of an average of 12% of the responses (varying from .82 to .92, single-measure ICC, two-way random, absolute agreement), administered tests were randomly distributed to be scored by individual raters. Finally, for each rater, stability of scoring was estimated. This ranged from .81 to 1.00. To provide more detailed information on the rating process, the interrater reliability per individual item is shown in Table 2. The low correlations of item 3 in all PAs indicate that raters may not have had a common understanding of the formulation of a hypothesis. On average, the rating process of a PA took 20 min per student.

Table 2 Intra-class correlations (α) of raters after training per item

		Skateboard	Bungee jump	Hot chocolate
Item 1	Formulate research question	.83	.74	.88
Item 2	Design experiment	.74	.69	.82
Item 3	Formulate hypothesis	.56	.61	.65
Item 4	Make a table	.86	.88	.81
Item 5	Make a graph	.89	.82	.90
Item 6	Interpret results	.92	.72	.57
Item 7	Extrapolate results	.88	.73	.75
Item 8	Draw conclusion	.86	.72	.85
Item 9	Formulate support for conclusion	.77	.60	.61
Item 10	Relate hypothesis to conclusion	.78	.67	.69
Item 11	Identify differences between plan and execution	.59	.59	.77
Item 12	Give suggestions to extend the experiment	.80	.83	.88
Item 13	Draw conclusion related to the context	.53	.51	.76
Item 14	Formulate learning gains	.88	.89	.83

Administration Procedure

Individual research assistants administered the PAs in groups of four to a maximum of eight students outside of the regular classroom. It took about 45 min to complete the test administration of one PA. Each research assistant received extensive training and followed detailed protocols for test administration. Each student completed three PAs on two different occasions with a time interval of 8 to 10 weeks. On the first occasion, all students completed the same PA (Skateboard). On the second occasion, administration of the two other PAs (Bungee jump, Hot chocolate) was randomly rotated. About half of the students completed the PA Bungee jump, while the other half completed Hot chocolate and vice versa. This rotation for the second occasion made it possible to determine whether both PAs map student performance in the same way, allowing us to conclude that they are equivalent diagnostic tools.

Findings

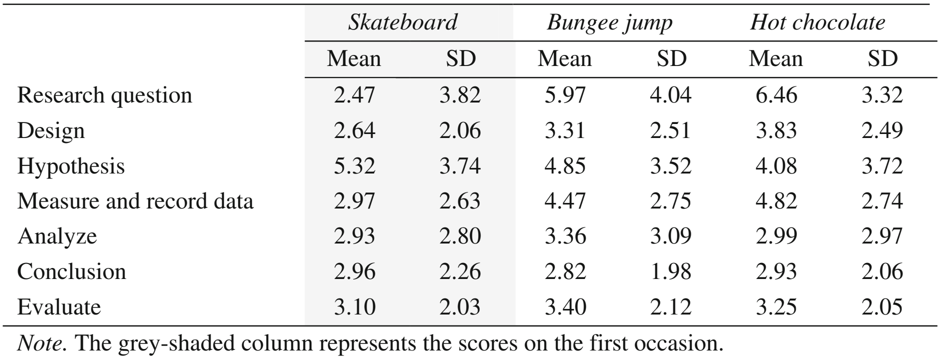
Descriptive Statistics

Table 3 presents the mean scores and standard deviations of the PAs per step. To facilitate comparison between the steps, the total scores for each step were converted into a standard scale ranging from 0 (lowest) to 10 (highest). In Table 4, the scores of Skateboard are shaded gray to emphasize that this PA was administered on a different occasion than the other two PAs. The mean scores on step level indicate differences between steps in terms of difficulty. For instance, designing an experiment seems in general to be more difficult than formulating a hypothesis. Differences are also visible between PAs. For example, in Bungee jump and Hot chocolate the mean scores for “measure and record” are higher than in Skateboard.

Overall, scores of the steps in all PAs show relatively low means indicating that the PAs were in general difficult for grades 5 and 6 students. The highest score is a 6.46 for formulating a research question. However, the large standard deviations reflect a high amount of variation within the sample.

In Table 4, the relationship between the PAs per individual step is displayed. Although significant, most correlations are small to medium with the exception of the step of “measure and record data” between Bungee jump and Hot chocolate which can be considered large

Table 3 Means and standard deviations on converted standard scales (0–10) for the different steps of the empirical cycle ($N = 403$)



	Skateboard		*Bungee jump*		*Hot chocolate*	
	Mean	SD	Mean	SD	Mean	SD
Research question	2.47	3.82	5.97	4.04	6.46	3.32
Design	2.64	2.06	3.31	2.51	3.83	2.49
Hypothesis	5.32	3.74	4.85	3.52	4.08	3.72
Measure and record data	2.97	2.63	4.47	2.75	4.82	2.74
Analyze	2.93	2.80	3.36	3.09	2.99	2.97
Conclusion	2.96	2.26	2.82	1.98	2.93	2.06
Evaluate	3.10	2.03	3.40	2.12	3.25	2.05

Note. The grey-shaded column represents the scores on the first occasion.

Note. The gray-shaded column represents the scores on the first occasion

Table 4 Correlations (Pearson's r) between the PAs per step of the empirical cycle ($N = 403$)

	Skateboard/Bungee jump	Skateboard/Hot chocolate	Bungee jump/ Hot chocolate
Research question	.15**	.25*	.24*
Design	.22*	.24*	.44*
Hypothesis	.06	.17**	.19*
Measure and record data	.37*	.33*	.83*
Analyze	.25*	.08	.24*
Conclusion	.32*	.16**	.34*
Evaluate	.29*	.29*	.39*

*$p < .001$; **$p < .05$—significant correlations (two tailed)

($r = .83$) (Cohen 1988). These moderate correlations may have been caused by task differences. Although the PAs were similar in structure and items, the topics between the PAs varied. As previously discussed, familiarity with the topic can influence application of skills considerably, resulting in variations between PAs within student performance.

Furthermore, these correlations between PAs as presented in Table 4, show that for most steps, correlations between Bungee jump and Hot chocolate are slightly larger than correlations of either of these PAs with Skateboard which was administered 8 weeks before the other two. Several reasons may account for these results. In the 8 weeks preceding administration of Bungee jump and Hot chocolate, about two third of the students had received lessons in which they had been performing small investigations similar to the PAs. In addition, all students had experience with the Skateboard experiment on the first testing occasion. As a result, students were more familiar with the testing format on the second occasion which may explain the difference in performance.

Response Patterns

Tables 5, 6, 7, 8, 9, 10, and 11 show in more detail how students performed in each step of the empirical cycle by presenting the response pattern of scores. In each table, the scores of Skateboard are shaded gray to emphasize that this PA was administered on a different occasion than the other two PAs. To demonstrate how the response patterns

Table 5 Response pattern of scores for the step of formulating a research question ($N = 403$)

	Performance assessment					
	Skateboard		*Bungee jump*		*Hot chocolate*	
score	*n*	%	*n*	%	*n*	%
0	271	67.2	100	24.8	47	11.7
1	65	16.1	125	31.0	191	47.4
2	67	16.6	178	44.2	165	40.9

Note. The gray-shaded column represents the scores on the first occasion

Table 6 Response pattern of scores for the step of designing an experiment ($N = 403$)

	Performance assessment					
	Skateboard		*Bungee jump*		*Hot chocolate*	
score	*n*	%	*n*	%	*n*	%
0	81	20.1	66	16.4	47	11.7
1	184	45.7	154	38.2	120	29.8
2	83	20.6	83	20.6	112	27.8
3	41	10.2	65	16.1	81	20.1
4	12	3.0	23	5.7	30	7.4
5	2	0.5	12	3.0	13	3.2

Note. The gray-shaded column represents the scores on the first occasion

may provide diagnostic information to teachers, the trends in the student responses per step will be discussed in more detail and illustrated with examples.

Formulating a Research Question

Scores presented in Table 5 for formulating a research question clearly show a shift from the majority of students scoring 0 points in the PA Skateboard to more than 75% of students scoring 1 or 2 points in PAs of the second occasion.

The research questions formulated by students awarded with 0 points in Skateboard were in general either unrelated to the goal of the experiment of finding a relationship between two variables ("What happens when the marble does not roll against the paper wedge?") or were impossible to investigate ("Why do Jake or Ying go faster?") (see Appendix 2 for the scoring rubric for formulating a research question). In the PAs on the second occasion more students were accurately able to formulate a research question which described the relationship between the two variables. For instance, "Does the rubber band stretch more when people are heavier?" in Bungee jump or "Does the amount of water influence the cooling rate?" in Hot chocolate.

Interestingly, for Hot chocolate, the research questions frequently addressed the issue of what makes hot drinks turn cold ("How does the drink cool faster: by blowing or just waiting?" or "Can a hot drink cool down in different ways?"). A

Table 7 Response pattern of scores for the step of formulating a hypothesis ($N = 403$)

	Performance assessment					
	Skateboard		*Bungee jump*		*Hot chocolate*	
score	*n*	%	*n*	%	*n*	%
0	100	24.8	106	26.3	155	38.5
1	177	43.9	203	50.4	167	41.4
2	126	31.3	94	23.3	81	20.1

Note. The gray-shaded column represents the scores on the first occasion

Table 8 Response pattern of scores for the step of measuring and recording data ($N = 403$)

	Performance assessment					
	Skateboard		*Bungee jump*		*Hot chocolate*	
score	*n*	%	*n*	%	*n*	%
0	109	27.0	51	12.7	44	10.9
1	69	17.1	37	9.2	29	7.2
2	77	19.1	66	16.4	63	15.6
3	47	11.7	75	18.6	63	15.6
4	53	13.2	66	16.4	73	18.1
5	28	6.9	52	12.9	69	17.1
6	17	4.2	50	12.4	54	13.4
7	3	0.7	6	1.5	8	2.0

Note. The gray-shaded column represents the scores on the first occasion

possible explanation is that students may have been more familiar with the topic of the cooling of hot drinks than with skateboarding or bungee jumping.

Designing an Experiment

Students' combined scores on the two items representing the step of designing an experiment (Table 6) are spread in the lower regions of scores.

Typically, low overall scores were mainly the result of students having failed to describe how they intended to communicate their results. Also, their descriptions were in general not very specific ("Attach the weights and see how far it stretches."), or they presented a design that did not relate to the research question ("I will see how fast the marble goes. We need cubes, a card and a ruler."), or they paid too much attention to details which were not relevant ("(1) Put the cube on one side; (2) Put the green paper on the other side; (3) Put the ruler on the cube; (4) Put the green paper on the card; (5) Roll the marble."). Designs awarded with higher scores were in general more extensive descriptions and included relevant details, such as the number of planned measurements ("Needed: 8 rings, clipboard, rubber band, a paper clip. First time measuring I put 4 rings on the paperclip, second time measuring 3 rings and

Table 9 Response pattern of scores for the step of analyzing data ($N = 403$)

	Performance assessment					
	Skateboard		*Bungee jump*		*Hot chocolate*	
score	*n*	%	*n*	%	*n*	%
0	153	38.0	141	35.0	165	40.9
1	86	21.3	78	19.4	68	16.9
2	118	29.3	112	27.8	109	27.0
3	34	8.4	48	11.9	48	11.9
4	12	3.0	24	6.0	13	3.2

Note. The gray-shaded column represents the scores on the first occasion

Table 10 Response pattern of scores for the step of formulating a conclusion ($N = 403$)

	Performance assessment					
	Skateboard		*Bungee jump*		*Hot chocolate*	
score	*n*	%	*n*	%	*n*	%
0	79	19.6	62	15.4	60	14.9
1	68	16.9	84	20.8	79	19.6
2	75	18.6	81	20.1	88	21.8
3	70	17.4	93	23.1	72	17.9
4	52	12.9	44	10.9	63	15.6
5	38	9.4	27	6.7	28	6.9
6	18	4.5	12	3.0	9	2.2
7	3	0.7	0	0	4	1.0
8	0	0	0	0	0	0

Note. The gray-shaded column represents the scores on the first occasion

third time 1 ring. Every time I measure how far it goes down. I note the results in a table.").

Formulating a Hypothesis

To support the formulation of a hypothesis, students were provided with a sentence starter: "I think that …, because …". As shown in Table 7, for each PA, the scores for formulating a hypothesis are spread in similar ways with most students scoring not higher than 1 point.

In general, most students were able to formulate a prediction and were awarded with 1 point, but they typically failed to substantiate their prediction ("I think the rubber band will stretch more with a heavier person, because the person is heavier") or (referring to the research question: "I think it is true, because the heavier, the more it stretches."). Students with full scores provided an explanation for their prediction, for instance "I think that the heavier thing goes down, because the rubber band needs more strength to hold the heavier thing up." In some instances, students managed to use more abstract concepts, such as "I think that it will stretch more, because more mass, more weight, more gravity." An interesting finding is the relatively low interrater reliability score (see Table 2) in each PA of this particular step.

Table 11 Response pattern of scores for the step of evaluating ($N = 403$)

	Performance assessment					
	Skateboard		*Bungee jump*		*Hot chocolate*	
score	*n*	%	*n*	%	*n*	%
0	74	18.4	49	12.2	57	14.1
1	75	18.6	94	23.3	89	22.1
2	121	30.0	114	28.3	120	29.8
3	101	25.1	95	23.6	96	23.8
4	31	7.7	42	10.4	36	8.9
5	1	0.2	7	1.7	4	1.0
6	0	0	2	0.5	1	0.2

Note. The gray-shaded column represents the scores on the first occasion

Apparently, the raters found it difficult to distinguish between a well-formulated hypothesis and a poorly formulated hypothesis.

Measuring and Recording Data

The step of measure and record data included making a table for noting the results and drawing a line graph. The pattern of the scores presented in Table 8 show that students performed better in the second occasion of PAswhen students in the experimental and control conditions had more experience with graphs. The response patterns further indicate that students differ considerably in their ability to measure and record data. For instance, around 12% of the students did not succeed at all in scoring points for this step, but the same proportion of students scored as many as 6 or 7 points.

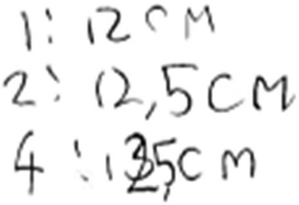
The most common approach of recording measurements was in a more or less structured way (see Fig. 1). Full credits were only awarded if a student recorded the data in a table indicating rows and columns (Fig. 2).

The graphs in each PA were pre-labeled to offer students some support. In Skateboard, many students made a bar graph instead of a line graph, indicating they did not have the specific knowledge on the concept of a line graph (Fig. 3). Furthermore, many students had difficulty to insert the units and the proper interval of units on both axes (Fig. 4) and using units with the right scaling resulting in graphs showing too little change (Fig. 5). However, some students were able to draw quite sophisticated graphs of their data, and one particular student even included a legend.

The examples provide information on how instruction on making graphs may improve scores for this particular step in the empirical cycle. For instance, students in the explicit condition received instruction on the purpose of different types of graphs and on how to decide on the units to put on the axis. To illustrate, Fig. 6 shows the progress in drawing graphs of one particular student in Skateboard and after 8 weeks in which the student had received instruction on drawing graphs in Bungee jump.

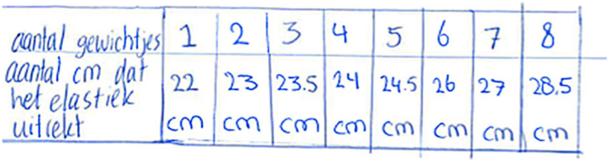
Analyzing Data

Finally, the steps of analyzing data, formulating a conclusion and evaluating show more or less similar patterns of scores between the three PAs. For analyzing data (see Table 9), full credits were only awarded if students explicitly mentioned the two variables and their relationship and if the conclusion was consistent with their own measurements. For instance: “If more rings are added, the rubber band gets longer.” or “As more rings are added, it gets heavier so the rubber band stretches more.” And for Skateboard: “The higher up the marble starts, the farther it rolls.”



1: 12CM
2: 12,5CM
4 : 13,5cm

Fig. 1 Noted results of measurements by a student in Skateboard



aantal gewichtjes	1	2	3	4	5	6	7	8
aantal cm dat het elastiek uitrekt	22 cm	23 cm	23.5 cm	24 cm	24.5 cm	26 cm	27 cm	28.5 cm

Fig. 2 Table awarded with full credits made by a student in Bungee jump

Students were in general able to describe the relationship between the two variables. However, although many students succeeded in describing the connection between variables, this relationship was not always supported by their own measuring and recording of data. Based on logical thinking, interdependence between variables was assumed by students, even if their own data did not match this reasoning. It was also regularly found that students indicated a relationship but failed to describe the relationship properly. For example, “More weights is more centimeters.” In Skateboard, students often referred to the speed of the marble as a reason for rolling farther away at the end of the ruler, such as “The higher it is, the faster it goes down.”

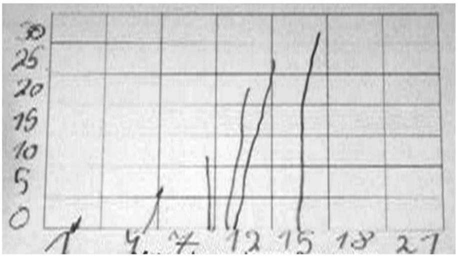
In extrapolating the results—also an element of analyzing results—much variation of students’ responses existed. No points were awarded if students simply mentioned an outcome. Students sometimes added an explanation but not substantiated with data. For instance, “I think 170 cm because it builds up much speed because of the length.” The following answer was awarded with full credits because the student had found a linear relationship between the two variables and used his own measurements: “2 meters far because if the slope is 20 cm it goes 10 cm far and two meters if 10 times as much as 20 cm, so 10 × 10 = 100 and that is 100 cm = 1 m.”

Formulating a Conclusion

The step of formulating a conclusion combined several aspects. First students were asked to give an answer to the research question and to support their answer with an explanation explicitly based on their own data. Then, students were required to relate the conclusion to the hypothesis as well as to the context of the PA. Table 10 shows how students’ scores were spread.

In all three PAs, approximately one third of the students were able to formulate a correct conclusion (“The higher you stand on the hill, the longer you can roll.”) but only an average of 9% of students achieved the maximum score of 2 points by supporting their conclusion with data.

Fig. 3 Example of a bar graph made by a student in Skateboard



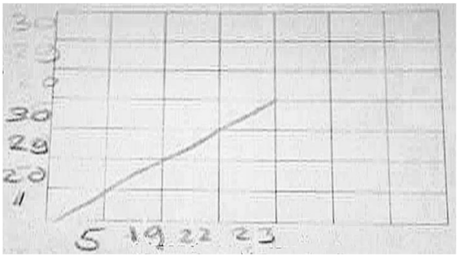


Fig. 4 Example of a graph with the units not inserted properly made by a student in Skateboard

When students were asked to relate the results to the specific context of the investigation, around 44% of the students in both Skateboard and Bungee jump, and 25% of students in Skateboard formulated a conclusion which related to the context (“With the heavier person the rubber band stretches more.”). In Bungee jump and Hot chocolate, few students were able to substantiate their conclusion by referring to the data. Remarkably however, 18% of students scored full credits for this particular item in Skateboard. During unstructured conversations after having completed the PAs on the second occasion, students told the researcher that they felt they were repeating their answers and therefore did not bother explaining their conclusion again, especially since it was their second and third PA.

Evaluating

The scores regarding the final evaluating step show that only three students obtained full credits (see Table 11).

In particular, an average of 9% of the students was able to give suggestions to extend the experiment by describing which relationship they would like to investigate (“To see whether the height on the hill also influences speed”). Most students receiving 1 point suggested additional experiments (see Table 1, item 12), but did not add an explanation (“I would do the experiment with longer distances”). Furthermore, about 80% of the students failed to formulate their learning gains resulting from the experiment in response to item 14 (see Table 1). For example, although students referred to having learned about the relationship between the variables which is considered a learning gain related to content, they did not answer the

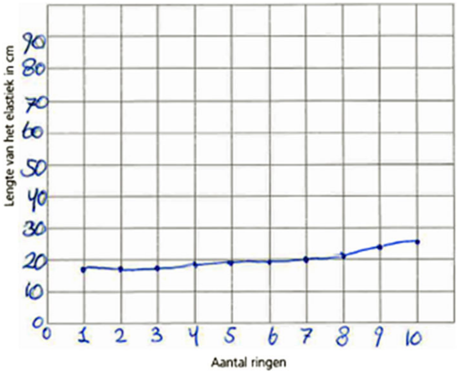


Fig. 5 Example of a graph with wrong scaling of units made by a student in Bungee jump

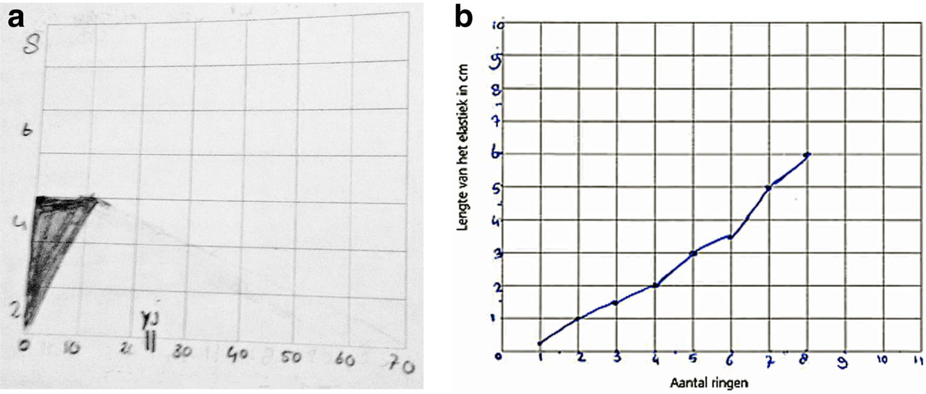


Fig. 6 Example of progress in drawing a graph of one of the students (Skateboard (**a**) and Bungee jump (**b**))

explicit question regarding what they had learned about performing an inquiry. However, the answers of the students who had received 1 or 2 points on item 14 were diverse and interesting. For instance, I have learned “that there are more ways to start an experiment,” “you get a better answer when you try it yourself,” “that measuring and being able to calculate is very important,” “how to perform an experiment by taking all steps,” “the reason for drawing a graph,” and variations of “that you have to work very precisely/neatly.”

Discussion

The present study aimed to answer the question whether a performance assessment can be used as a diagnostic tool to evaluate students’ progress and to guide instruction in science classroom practice. Findings show that PAs have potential as a diagnostic tool for monitoring students’ performance of skills, hence adjusting instruction and activities to enhance learning. In particular, the structuring of PAs by assigning items to the different steps of the empirical cycle, combined with the extensive descriptions of performance expectations for each item, has shown to be useful in evaluating students’ responses.

This structured approach makes it possible for teachers to analyze the responses of students on various levels and use the findings to adjust their instruction. For instance, the means per step provide information on how students perform as a group at the classroom level. Response patterns per step indicate that there is considerable variation, indicating that the measurements are suitable for mapping individual differences. In particular, the response patterns of the steps reveal where students show particular difficulty and in which steps students perform more successfully compared with the average classroom performance. This information will help the teacher to obtain an overall picture of how students are progressing and subsequently adjust instruction as well as make informed decisions regarding the choice of science activities. Finally, looking at students’ responses in more detail provides insight into the common mistakes the students make and reveals to some extent students’ thinking processes (Davey et al. 2015). As a result, teachers are able to not only adjust their instruction to remedy shortcomings but also give specific feedback to individual students (Black et al. 2004; Harlen 1999).

Similarly, the rubric may have additional value for implementing PAs as a diagnostic tool. The results of analyzing interrater reliability and consistency of scoring suggest that the scoring rubric can be used effectively by trained raters. In particular, the high consistency of scoring shows that the raters were able to apply the rubric to score students' written answers reliably. This implies that for teachers, it is possible to assess students' answers to different PAs consistently over time by using the rubric.

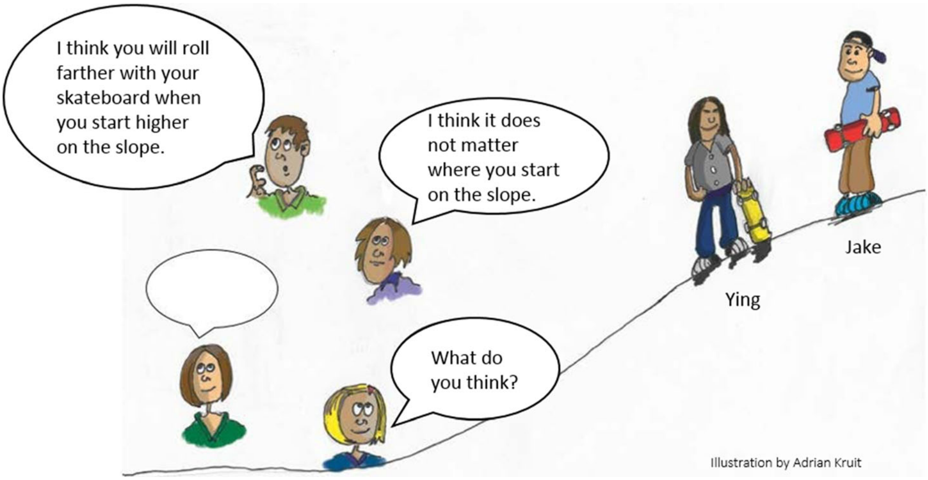
In addition, the rubric can play an important role in the professional development of science teachers. Because of the extensive description of the learning objectives, of which teachers do not always have a clear understanding (Aschbacher and Alonzo 2006), and the different levels of proficiency, teachers may become more explicitly aware of the learning objectives of scientific inquiry, while at the same time gaining better understanding of the skills they are scoring. Emphasis on scoring is particularly important, since in the Netherlands only in 16% of primary schools (grades 4–6), the progress of students' learning of science is monitored by teachers (Inspectorate of Education 2015).

The diagnostic value of PAs may also prove to have more value than just of an assessment tool. Davey et al. (2015, p. 9) state that "A good assessment task is a good teaching task and vice versa," which stresses the importance of alignment between curriculum and assessment. Because of the structuring of the PAs in steps of the empirical cycle, a PA can be split and administered in more than one lesson as a component of a regular science lesson. In this way, teachers can monitor performance through the workbooks of students which provides the opportunity to give feedback during the course of several lessons (Harlen 1999). Furthermore, it addresses the problem of the limited time assigned for science instruction in most primary schools (Martin et al. 2012; National Academies of Sciences, Engineering, and Medicine 2015). Finally, adding the rubric as part of the instruction may also create awareness and understanding of the science skills and students' own learning process, which is perhaps the most important purpose of formative assessment (Harlen 1999).

There is one limitation that should be addressed when implementing the PAs. The response patterns of the scores vary between the PAs as revealed by the low and moderate correlations between PAs per step. This may be attributed to the different topics used for the PAs. Students may find one topic more interesting or less difficult than another. Teachers may choose PAs with topics with which students are familiar (either by own experience or being taught about the content of the PA topic) or in which they are interested. In addition, inconsistencies in rating may have influenced variation between steps of the PAs (Ruiz-Primo et al. 1993). As discussed above, although interrater reliability was high on total scores as well as on most individual items, few items proved more difficult to score such as formulating a hypothesis. This suggests that although formatively evaluating on step level is useful for teachers to monitor (a group of) students, the scores on step level may not be reliable enough for summative assessment. For reliable summative assessment, more items per step should be included or more PAs with a range of different topics should be used.

In summary, this study shows that a PA structured according to the steps of the empirical cycle, is a useful tool to inform teachers on students' science skills at a detailed level. It does not require intensive preparation to administer in a science classroom and fairly simple materials can be used. Implementation of the PAs need not necessarily be limited to grades 5 and 6 but may also be used for students in grades 7 or higher. Professional development of teachers should address the learning objectives for scientific inquiry and how to use students' responses for evaluating these responses. Future research will have to determine to what extent science teachers will be willing and able to implement the PAs as a diagnostic tool in their classroom.

Appendix 1. Performance assessment Skateboard


I think you will roll farther with your skateboard when you start higher on the slope.
I think it does not matter where you start on the slope.
What do you think?
Ying
Jake
Illustration by Adrian Kruit

Jake and Ying want to roll down the hill with their skateboards. Each boy starts at a different height. Jake thinks that he will roll farther than Ying at the bottom of the hill. Ying thinks that he will go farther than Jake. You are going to investigate this. It is not possible to skateboard in the classroom, so the experiment will be performed on your school desk.

First, you will be required to complete the following assignments:

Task 1:

Can you formulate a question pertaining to your experiment for which you would like to find the answer? Write down your question.*

Task 2:

You will carry out an experiment in which you will try to find an answer to the following question:

Will you roll farther when you start higher up on the slope?

There are materials on the table. You will let the marble (= skateboard) roll from the ruler (= slope). On another table, you will see an example of the set-up you are going to use. Build the set-up with your own materials.

Now you are going to make a design for your experiment. Describe in steps how you will perform your experiment. For example, think about what you will measure, how and how often you will measure it. Also, describe how you will record the results of your experiment.

Task 3:

What do you think the answer will be for the research question: *Will you roll farther when you start higher up on the slope?*

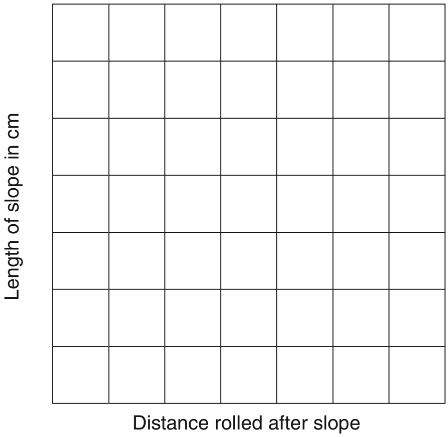
I think that ……………………………………….., because …………………………………………………

Task 4:

You will execute your experiment following your own design in Task 2. The purpose is to find an answer to the research question: *Will you roll farther when you start higher up on the slope?*

a) Record the results in the space below. Think about how you will write down the results.

b) Make a line graph of your results:



Task 5:

You can answer the following questions by looking at your results and your graph.

a) Describe how the distance the marble rolls on the slope relates to how far the marble continues to roll beyond the slope.

b) Suppose that you could make a slope that is 2 meters long. How far would the marble be able to roll beyond the slope? Explain your answer.

Task 6:

a) Reread the research question for task 2. What will your answer be now?

b) Which reasons can you give to show that your answer is correct?

Task 7:

a) In Task 3 you described the results you expected. Now that you have done the experiment, which similarity and/or difference is there between what you expected and what you found in the investigation? Explain your answer.

In Task 4 you designed your experiment. The following three questions are about the *execution* of your investigation.

b) What – if anything - did you end up doing differently than was described in your design?

c) If you did anything differently, what was your reason for doing so?

d) If you carried out your design according to your plan, what would you do differently next time round in order to improve upon your experiment? Explain your answer.

e) Perhaps your curiosity was piqued after having performed this experiment. Suppose that you could use different (or more) materials and had more time. What would you investigate or change about your experiment regarding skateboarding?

f) You investigated the distance travelled on the slope and the distance the marble continued to roll beyond the slope. What can you now say about the rolling of Jake and Ying's skateboards?

g) What did you learn about doing an experiment?

* In the actual assessment, each question was followed by dotted lines for students to write their responses on (Kruit et al. 2018b).

Appendix 2

Table 12 Example of scoring rubric for the skill 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. 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 substracted.
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?	
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 farther 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?	

References

Abrahams, I., & Reiss, M. J. (2015). The assessment of practical skills. *School Science Review., 96*(357), 40–44.

Ahlbrand, W., Green, W., Grogg, J., Gould, O., & Winnett, D. A. (1993). *Science performance assessment handbook*. Edwardsville: Illinois Science Teachers Association.

Alonzo, A. C., & Aschbacher, P. R. (2004, April). Value-added? Long assessment of students' scientific inquiry skills. *Proceedings of assessment for reform-based science teaching and learning*. Symposium conducted at the annual meeting of the AERA, San Diego, CA.

Aschbacher, P., & Alonzo, A. (2006). Examining the utility of elementary science notebooks for formative assessment purposes. *Educational Assessment, 11*(3–4), 179–203.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: benchmarks and surrogates. *International Journal of Educational Research, 21*(3), 279–298.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). *Working inside the black box: assessment for learning in the classroom*. London: GL Assessment.

Black, P., & Wiliam, D. (2003). 'In praise of educational research': formative assessment. *British Educational Research Journal, 29*(5), 623–637.

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

Britton, E. D., & Schneider, S. A. (2014). Large-scale assessments in science education. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (Vol. 2, pp. 791–808). Abingdon: Routledge.

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: Lawrence Earlbaum Associates.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (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., Bransford, J. D., & Pellegrino, J. W. (1999). *How people learn: bridging research and practice*. Washington, DC: National Academies Press.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking science to school: learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

Eberbach, C., & Crowley, K. (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., Miller, P. H., & Miller, S. A. (1993). *Cognitive development*. New Jersey: Prentice-Hall.

Germann, P. J., & Aram, R. J. (1996). Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of Research in Science Teaching, 33*(7), 773–798.

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., & Duggan, S. (1995). *Investigative work in the science curriculum*. Buckingham: Open University Press.

Gott, R., & Duggan, S. (2002). Problems with the assessment of performance in practical science: which way now? *Cambridge Journal of Education, 32*(2), 183–201.

Gott, R., & Murphy, P. (1987). *Assessing investigation at ages 13 and 15: assessment of performance unit science report for teachers: 9*. 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. (1999). Purposes and procedures for assessing science process skills. *Assessment in Education: Principles, Policy & Practice, 6*(1), 129–144.

Harlen, W., & Qualter, A. (2009). *The teaching of science in primary schools*. Abingdon: Routledge.

Harlen, W., Bell, D., Devés, R., Dyasi, H., de la Garza, G. F., Léna, P., & Yu, W. (2012). *Developing policy, principles and practice in primary school science assessment*. London: Nuffield Foundation.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I., Gonzalez, E. J., & Orpwood, G. (1997). Performance assessment: IEA's Third International Mathematics and Science Study (TIMSS). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Inspectorate of Education. (2015). *Wereldoriëntatie. De stand van zaken in het onderwijs. [World orientation. The present state of education]*. Utrecht: Inspectie van het Onderwijs.

Kind, P. M. (1999). Performance assessment in science—what are we measuring? *Studies in Educational Evaluation, 25*(3), 179–194.

Klassen, S. (2006). Contextual assessment in science education: background, issues, and policy. *Science Education, 90*(5), 820–851.

Kruit, P. M., Oostdam, R. J., van den Berg, E., & Schuitema, J. A. (2018a). Assessing students' ability in performing scientific inquiry: instruments for measuring science skills in primary education. *Research in Science & Technological Education*. https://doi.org/10.1080/02635143.2017.1421530.

Kruit, P. M., Oostdam, R. J., van den Berg, E., & Schuitema, J. A. (2018b). Effects of explicit instruction on the acquisition of students' science inquiry skills in grades 5 and 6 of primary education. *International Journal of Science Education, 40*(4), 421–441.

Kuhn, D. (1997). Constraints or guideposts? Developmental psychology and science education. *Review of Educational Research, 67*(1), 141–150.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction, 9*(4), 285–327.

Lederman, N., & Lederman, J. (2014). Research on teaching and learning of nature of science. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (Vol. 2, pp. 600–620). Abingdon: Routledge.

Loxley, P., Dawes, L., Nicholls, L., & Dore, B. (2013). *Teaching primary science: promoting enjoyment and developing understanding*. Harlow, England: Pearson Education.

Martin, M. O., Mullis, I. V., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (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.

Martin, M. O., Mullis, I. V., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill: TIMSS & PEARLS International Study Center, Boston College.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Michaels, S., Shouse, A. W., & Schweingruber, H. A. (2007). *Ready, set, science!: putting research to work in K-8 science classrooms*. Washington, DC: National Academies Press.

Millar, R., & Driver, R. (1987). Beyond processes. *Studies in Science Education, 14*(1), 33–62.

Moseley, D., Baumfield, V., Elliott, J., Gregson, M., Higgins, S., Miller, J., & Newton, D. P. (2005). *Frameworks for thinking: a handbook for teaching and learning*. Cambridge: Cambridge University Press.

National Academies of Sciences, Engineering, and Medicine (2015). *Science teachers learning: enhancing opportunities, creating supportive contexts. Committee on Strengthening Science Education through a Teacher Learning Continuum. Board on Science Education and Teacher Advisory Council, Division of Behavioral and Social Science and Education*. Washington, DC: National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

National Research Council. (2014). In Committee on Developing Assessments of Science Proficiency in K-12, Board on Testing and Assessment and Board on Science Education, J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty (Eds.), *Developing Assessments for the Next Generation Science Standards*. Washington, DC: National Academies 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.

NGSS Lead States. (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.

OECD. (2017). *PISA for development assessment and analytical framework*. Paris: OECD Publishing.

Osborne, J. (2015). Practical work in science: misunderstood and badly used? *School Science Review, 96*(357), 16–24.

Pellegrino, J. W. (2012). Assessment of science learning: living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831–841.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: a systems approach to progress. *Psicología Educativa, 20*(2), 65–77.

Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., et al. (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.

Roth, K. J. (2014). Elementary science teaching. In N. G. Ledermann & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. 2, pp. 361–393). New York: Routledge.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41–53.

Schilling, M., Hargreaves, L., Harlen, W., & Russell, T. (1990). *Assessing science in the primary classroom: written tasks*. London: Paul Chapman Publishing.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347–362.

Shavelson, R. J., Solano-Flores, G., & Ruiz-Primo, M. A. (1998). Toward a science performance assessment technology. *Evaluation and Program Planning, 21*(2), 171–184.

Schraw, G., Crippen, K. J., & Hartley, K. (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.

Solano-Flores, G., Javanovic, J., Shavelson, R. J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21*(3), 293–315.

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The effects of content, format, and inquiry level on science performance assessment scores. *Applied Measurement in Education, 13*(2), 139–160.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: what will it make it work? In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education, 123*(1), 39–56.

Zohar, A., & Dori, Y. J. (2003). Higher order thinking skills and low-achieving students: are they mutually exclusive? *The Journal of the Learning Sciences, 12*(2), 145–181.