‘Would you dare to jump?’ Fostering a scientific approach to secondary physics inquiry

Pols, C. F.J.; Dekkers, P. J.J.M.; de Vries, M. J.

DOI
10.1080/09500693.2022.2083251

Publication date
2022

Document Version
Final published version

Published in
International Journal of Science Education

International Journal of Science Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsed20

'*Would you dare to jump?*' Fostering a scientific approach to secondary physics inquiry

C.F.J. Pols, P.J.J.M. Dekkers & M.J. de Vries

To cite this article: C.F.J. Pols, P.J.J.M. Dekkers & M.J. de Vries (2022): '*Would you dare to jump?*' Fostering a scientific approach to secondary physics inquiry, International Journal of Science Education, DOI: 10.1080/09500693.2022.2083251

To link to this article: https://doi.org/10.1080/09500693.2022.2083251

Published online: 13 Jun 2022.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION
https://doi.org/10.1080/09500693.2022.2083251

Routledge
Taylor & Francis Group

'*Would you dare to jump?*' Fostering a scientific approach to secondary physics inquiry

C.F.J. Pols, P.J.J.M. Dekkers and M.J. de Vries

Science Education and Communication, University of Technology Delft, Delft, The Netherlands

ABSTRACT

Secondary school students often only use the rules for doing scientific inquiry when prompted, as if they fail to see the *point* of doing so. This qualitative design study explores conditions to address this problem in school science inquiry. Dutch students ($N = 22$, aged 14–15) repeatedly consider the quality of their work: in a conventional, guided inquiry approach; by evaluating their conclusion in terms of the contextual purpose of the investigation; as consumers of knowledge facing the (hypothetical) risk of applying the findings in the real world. By gauging students' confidence in the inquiry's trustworthiness, we established that, while each confrontation instigated some students to (re)consider the quality of their inquiry, the final stage had the greatest impact. Students came to see that finding trustworthy results is essential, requiring scientific standards. The scientific quality of their inquiries was described, weaknesses identified and compared with the improvements students themselves proposed for their inquiries. While the improvemens were expressed in non-specific terms these align with a scientific perspective. Students now *wanted* to find trustworthy answers by exploiting scientific standards. In enabling students to engage successfully in basic scientific inquiry, finding ways to establish students' mental readiness for attending to the quality of their scientific claims, and of personalised scientific criteria for their assessment, is indispensable.

ARTICLE HISTORY
Received 26 July 2021
Accepted 24 May 2022

KEYWORDS
Scientific inquiry; argumentation; practical work; Concepts of Evidence; physics

Introduction

Practical work refers to activities in which students manipulate instruments and materials to answer a research question (Millar et al., 1999). It is frequently used to achieve two broad aims in science education: (1) to help students develop a proper understanding of the relation between scientific theory and practice, and (2) to become competent in conducting their own scientific research (Abrahams, 2005; Hodson, 2014; Hofstein, 2017; Hofstein & Kind, 2012; Millar, 2004; Millar et al., 1999). This paper focuses on the second of these aims. Despite many decades of research

CONTACT C.F.J. Pols c.f.j.pols@tudelft.nl Science Education and Communication, University of Technology Delft, Lorentzweg 1, Room F232, Delft 2628 CJ, The Netherlands

and development, practical work in most of today's classrooms still involves students doing no more than following up on detailed instructions (Abrahams & Millar, 2008; Holmes & Wieman, 2016, 2018; Wieman, 2015). When instructed to do so, the students repeat measurements sufficiently often, calculate averages correctly, apply appropriate instruments, use suitable tables and graphs, etcetera. But as soon as we stop telling them what to do, they stop doing so, and are unable to find valid and reliable answers by themselves (Millar, 2004). In other words, we have been unable to use practical work effectively to enable students to engage in basic scientific inquiry independently (Abrahams, 2011; Abrahams & Millar, 2008; Hofstein, 2017; Hofstein & Kind, 2012; Hofstein & Lunetta, 2004; Lunetta et al., 2007).

As many scholars before us we believe practical work aimed at teaching students how to engage in basic scientific inquiry often lacks opportunity for students to learn from their own (methodological) mistakes and fails to provide a sense of (scientific) purpose (Hodson, 2014; Holmes & Wieman, 2016, 2018; Wieman, 2015). Each practical activity tends to be a standalone event rather than an integrated part of a coherent approach to developing understanding of and competence in scientific inquiry. Disappointing learning outcomes regarding practical work may in part be caused also by a lack of relevance for students. In absence of any practical importance of their investigations it is unlikely they will value the quality of the outcome or invest much effort in obtaining it. Indeed students often carry out measurements rapidly with insufficient attention to care and precision (Millar et al., 1999) resulting in unreliable data and superficial, incomplete conclusions (Kanari & Millar, 2004; Pols et al., 2021).

Practical work, according to the literature, should be made more 'open', allowing students to make their own choices (Glaesser et al., 2009; Hodson, 2014; Hofstein & Kind, 2012; Holmes & Wieman, 2016, 2018; Zion & Mendelovici, 2012). Indeed, in our personal professional experience, when we make practical work more open we see that they tend to make choices that optimise their work. Unfortunately, students usually optimise it in terms of the time and effort that they invest, not in terms of the scientific quality of the answer to the research question (Pols et al., 2021).

This paper is based on the assumption that before we can expect students to make desirable choices in inquiry, we will have to teach them the value of that *scientific quality*. We explore an educational design aimed at developing in students the understanding that in scientific inquiry, one seeks the best possible answer in the given circumstances (Lipton, 2003). Our intervention aims to develop in students an *intent* to obtain a scientifically adequate answer, and an *understanding* of what makes it scientifically adequate. Of course, one of the problems we will have to solve is that students do not yet know what we mean, in a scientific sense, by 'the best possible answer in the given circumstances'. After the activity, we do not expect them to have become proficient researchers, but to have developed a mindset that is directed at making choices that optimise the *quality* of the answer to the research. Personal reasons and intentions for producing scientifically sound research may contribute to students accepting and applying the taught rules and practices in more independent physics inquiry and may motivate them to further develop their understanding of these rules and practices (Kortland, 2007). This is a first step in addressing the challenge identified by Hofstein (2017): *to help learners take control of their own learning in the search for understanding while providing opportunities that encourage them to ask questions, suggest hypotheses, and design*

investigations. We will present the research questions after the educational design, below, since their specific contents depend on it.

Theoretical framework

'The quality of the answer to the research question' is analysed in terms of a theoretical model that describes the different types of knowledge applied in scientific inquiry. Design choices regarding the 'openness' and contextualisation are clarified next.

A model of the knowledge applied in practical work – PACKS

Practical work should be a minds-on activity characterised by students' use of their *Procedural and Conceptual Knowledge in Science (PACKS)* (Millar et al., 1994). Figure 1 presents the PACKS model and the different types of knowledge (**A**–**D**) that influence researchers' decisions in the different stages of an inquiry (Millar et al., 1994). The model distinguishes knowledge of (**A**) the nature and purpose of the inquiry, (**B**) relevant content, (**C**) required manipulative skills and (**D**) evaluating scientific evidence. Consideration of the quality of the research involves, in the first place, application of type **D** knowledge, comprising of awareness and use of criteria involved in the construction and evaluation of scientific evidence. These criteria include, *i.a.*, an operationalisation of the *Concepts of Evidence (CoE)*. These are concepts such as fair test, experimenter bias, range, median, precision and measurement uncertainty that underpin the more abstract concepts of reliability and validity (Gott et al., 2003; Gott & Duggan, 1996). Using a scientific approach in practical work entails the conscious and adequate use of this type of knowledge in finding and evaluating answers to the question: *At this point, what needs to be done to achieve the best possible result in this investigation in the given circumstances?* 'Best possible result', that is, in terms of the scientific goal of describing, explaining and predicting events and phenomena as precisely and accurately as possible.

Understandings of Evidence

Rather than evaluating the presence of isolated concepts, we proposed to consider collections of loosely interrelated CoE that constitute overlapping '*Understandings of Evidence*' (UoE) (Pols et al., 2022). UoE express properties of the evidential information at a

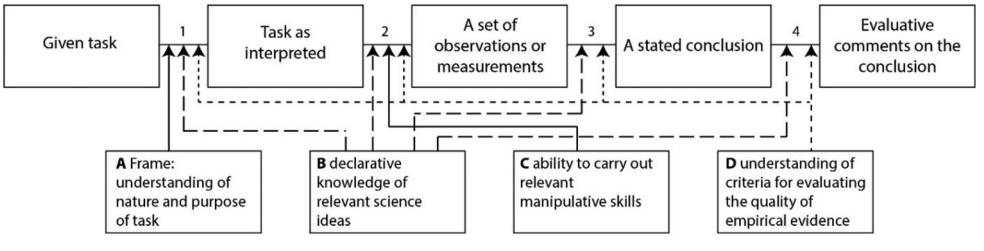

Figure 1. The procedural and conceptual knowledge in science model (Millar et al., 1994) illustrates how different types of knowledge, on the left, influence decisions made at various stages of an inquiry.

particular stage, or procedures for constructing that information, as well as prescriptions for enhancing or assessing informational quality. The UoE delineate knowledge a researcher has and applies in constructing an optimally reliable and valid inquiry. They are the common understandings by which researchers evaluate and judge the quality of their own research and that of others, the norms and standards they use to determine how well the empirical data support the researcher's claims.

An Assessment Rubric for Physics Inquiry (ARPI) was constructed and validated by the authors (Pols et al., 2022) that allows for assessment of a student's UoE based on his or her observable inquiry actions and research report. The instrument distinguishes 19 UoE distributed across six phases of inquiry: (1) Asking questions, (2) Design, (3) Methods & procedures, (4) Analysis, (5) Conclusion and evaluation, (6) Peer review. For each UoE, indicators for the lowest, intermediate and highest levels on a five point scale are provided, see Table 1, where levels in between are assigned when a student outperforms the lower level but not fully attains the higher level. Depending on the openness of the inquiry, the precise task and the specific learning goals, specific (clusters of) UoE can be selected for assessment purposes. For instance, in structured inquiry (Table 2), only ARPI's clusters (4)-(6) can be assessed as students are using a given research question and method.

ARPI is used here to evaluate the scientific quality of students' inquiry approach in the given tasks, based on the actions, decisions and justifications found in their research reports. It is also used to establish the scientific quality of the ideas students forward to enhance the quality they ascribe to their own work.

Guided inquiry

While developing understandings of scientific inquiry requires that students (be given the opportunity to) take agency and learn from the consequences (Hodson, 1992), inexperienced students conversely need support and structure. In terms of student input and choice, *guided inquiry* offers a balance (Banchi & Bell, 2008; Tamir, 1991) that is appropriate in this study. Table 2 shows that the research question is posed by the teacher but the answer is unknown by the students beforehand and they decide on the procedure. Students will all attempt to answer the same research question. However, depending on their ideas about evidence and their understanding of what constitutes 'good science' (Gott & Duggan, 1996), they will make different decisions. Students will thus differ in how they answer the research question, and in the scientific quality of that answer. Since all students can consider the quality of their own and each other's work in terms of the same scientific purpose, this quality can become the focus of attention rather than the details of subject matter, the experimental setup or the data analysis.

Context-based approach

A context-based approach is advocated in various curricula including the Dutch physics curriculum (Bennett et al., 2007; de Putter-Smits, 2012; Netherlands Institute for Curriculum Development, 2016) as intrinsically more authentic, stimulating and interesting. Students are assumed to put more effort into learning content that is perceived as

Table 1. Illustrative excerpts of the Assessment Rubric for Physics Inquiry (Pols et al., 2022). Five levels of competence are distinguished for each Understanding of Evidence described on the left. Descriptors for lowest, intermediate and highest levels are specified, where intermediate levels can be assigned when a student outperforms the lower level though not yet fully attains the higher level.

UoE		Level of competence		
Phase	The researcher understands that:	0	2	4
Method & procedure	8: measured values will show inherent variation and the reliability of data must be optimised, requiring repeated measurements.	Collects too few repeated measurements without substantiation or consideration of the quality of the dataset. Does not consider collecting further data at any stage.	Repeats measurements a fixed but sufficient number of times without substantiation in terms of the quality of the dataset. Considers collecting additional data only in retrospect, as a recommendation.	Substantiates the required number of repeated measurements based on the spread in the data and the required reliability. Considers collecting alternative, additional data and collects these if appropriate.
Analysis	12: Data require appropriate methods for analysing and describing them.	Chooses inappropriate data representations.	Chooses suitable but not optimal data representations to establish a pattern.	Makes use of appropriate data representations, clearly revealing the pattern and features in the data.
	13: An optimally informative answer to the research question requires a description of relationships in as much detail as possible. Quantitative descriptions are more detailed than qualitative ones.	Expresses relationships in a qualitative sense only.	Describes patterns correctly but misses some details of features or mathematical properties in relationships.	Describes patterns in appropriate detail. Specifies a mathematical expression or describes the quantitative relationship of the dataset if possible.

Table 2. Tamir (1991) distinguishes four levels of inquiry, depending on the information provided to the student. Guided inquiry balances the teacher's support with students' independency to organise the research as they see fit.

Inquiry type	Question/problem	Method/procedures	Conclusion/solution
Confirmation	Given	Given	Given
Structured	Given	Given	open
Guided	Given	Open	open
Open	Open	Open	open

relevant because of its context (Kortland, 2007). However, if a context merely serves to teach difficult concepts, students quickly lose interest (Kortland, 2007; Lijnse, 2014, p. 157) and forget the context (Molyneux-Hodgson et al., 1999).

The relevance of the context in this intervention rests in a CoE called the *practicality of consequences* (Gott & Duggan, 2003, CoE 87), i.e. the practical implications of applying the findings of an inquiry. While it rarely plays a role in conventional practical work, we use it to try and entice students to demand, without having to be told to do so, the highest possible standards of validity and reliability of the evidence. This specific CoE contributes to raising awareness of the nature and purpose of the given task (PACKS knowledge type **A**). Its use may help in *holding students accountable for the quality of their results* (Duschl, 2000) and scaffold students' use of a scientific attitude towards producing sound research (Ntombela, 1999, p. 127).

Method

This section presents the research design, and then describes the participants and the Dutch educational context. Next the educational design, research questions, data collection and analysis are addressed.

Research design

Informed by the literature on teaching inquiry in science education, practical work and context-based approaches, we developed an intervention consisting of three stages. Each stage has a different approach to fostering students' consideration of the quality of their answer to the research question. In the first stage, a lesson of 50 min, the purpose of the investigation is clarified and a conventional, guided inquiry approach followed. The next stage involves a homework assignment in which the context is invoked so as to ask students to report about their findings to a hypothetical outsider. One week later, the students are asked in the third stage, also a 50 min lesson, to consider their results as *consumers* of the research outcomes rather than as its *producers*. We study whether, when and how the students' consideration of the quality of their inquiry changed and how it depends on the characteristics of these specific stages. To do so, a qualitative small-scale developmental design study in an authentic setting was chosen. This design, with a high degree of ecological validity (Brewer, 2000), allows for closely monitoring students' approaches to the inquiry through evaluation of the written accounts of their work, analysis of recorded discussions and of their self-evaluation forms.

Participants & educational context

The study was conducted in the spring of 2019 in an intact Grade 9 class of an urban school in the Netherlands. Participation was mandatory and graded, but while work of higher quality did earn a higher grade, attending and handing in the work sufficed to earn a passing grade.

The teacher, also the first author of this paper, had 9 years of teaching experience in physics at secondary school. Well aware of the challenges involved he had conducted several in-service and conference workshops on practical work and teaching scientific inquiry (Pols, 2021b). As advocated in the literature, teachers' research of their own practice is an authentic way to study 'what goes in the school laboratory' (Hodson, 1990; Hofstein, 2017) and the students' behaviour and constructed perceptions and understandings (Hofstein and Kind, 2012). It has the potential to close the research-practice gap (Bakx et al., 2016).

Convenience sampling was used as the intervention was designed and carried out by the regular teacher of the 23 students. The students, aged 14-15, were in their last year of lower secondary education, physics still being a mandatory subject. While broad guidelines are provided as to content and level (Ottevanger et al., 2014; Spek & Rodenboog, 2011), in the absence of a national exam program for lower secondary school, attainment levels cannot be precisely defined. Although it is meant to develop scientific literacy, this compulsory part of science education does not actually provide students with proficiency in independent inquiry. The study of Pols et al. (2021), carried out in the same population, concludes that students rely on the teacher's input rather than their own resources when it comes to producing scientifically sound research. If the students in the current study have some (implicit) understandings of inquiry, these result from closed 'cookbook' experiments that tell students precisely what to do.

Educational design

In the first stage, students watched a spectacular scene from a popular film, where pirates swing on thick ropes from one sailing ship to another while sharp objects fly and serious explosions go off all around (Bruckheimer, 2007). They were tasked to help the stunt coordinator plan a novel film stunt that should be spectacular but safe for the stunt people. They were to gather the required information from studying a pendulum, as a model for swinging on ropes between ships. The class worked together in identifying factors that might influence the 'swing time' and small teams were formed to each investigate one of these. No further guidance was given in terms of procedure or required answers, but it was emphasised that the pirate is to arrive shortly after a big blast. Arriving too early would be dangerous, too late would be insufficiently spectacular and require expensive retaking of the scene.

The teacher's role during the first part of the intervention was modest. He was to explain the task, emphasising that the stunt was to be filmed in a single take. Students were then expected to devise the experiment as they see fit. If students had questions related to the given task, to the physics involved or to the use of (more advanced) research methods and instruments (knowledge types **A–C**), these were to be answered directly so as to reduce the chance of cognitive overload. This would allow students to

focus on knowledge type **D** only (Johnstone & Wham, 1982; van den Berg, 2013). If students had questions addressing knowledge type **D**, or if the teacher observed errors in, e.g. controlling variables, these issues were to be discussed on the spot.

Students had access, in principle, to more sophisticated measuring apparatus available in the school lab, to measuring techniques involving, e.g. their mobile phones and to internet sources. The teacher was to provide assistance with use of these options but only at students' request. Help and materials were provided only if students expressed, of their own accord, dissatisfaction with the quality of their evidence. Therefore, if an optimal quality of evidence was not obtained, we can attribute this to deficiencies in their (application of) type **D** knowledge. It cannot be explained by a lack of type **C** knowledge about measuring apparatus. Since no attention was paid to the match of students' findings with the accepted description of the physical pendulum at any stage, no interference from type **B** knowledge about physics content is involved either. All measurements and inferences are accepted as given.

Apart from the use of a film clip, this approach so far is conventional. Since inexperienced students tend to be brief and superficial in their construction, justification and evaluation of conclusions in inquiry, and the intervention so far does not affect this, no serious consideration of the quality of the answer to the research question was expected. A conventional practical would end here, with a brief lab report on what factors affect the period of the swing (and possibly a teacher explanation of the appropriate formula).

The intervention, however, proceeded with a homework assignment, referred to as the second stage of the intervention. It required student teams to write a letter to the stunt coordinator to explain what they investigated and found, and whether they thought their results were useful for designing the new stunt. Invoking the context was to provide students with more tangible reasons to elaborate on the quality of their answers than filling in a lab report does. It was meant to stimulate taking accountability for conclusions, justifying research actions and discussing the trustworthiness of the findings. Since still no particular personal relevance was attached to the outcome of the inquiry, however, we expected the impact to be limited, and most students to perform the task in the usual way – compliant but with minimum effort.

Teams submitted their letters online, enabling the teacher to establish the students' reports as input for the reflective evaluation of the inquiry in the next lesson, the third stage of the intervention. In this evaluative stage of the inquiry, the students' perspective of the context was meant to become that of the *consumers* of the knowledge produced. They were asked to evaluate their inquiry from the perspective of the stunt(wo)man: '*would you dare to jump, if the stunt was based on the information you have provided?*' The *practicality of consequences* for students is meant to change from 'being judged on my report' to 'risking my life' (or rather, imagining what the implications are if the research findings are actually used). Much depended on whether students were prepared to take their assigned role seriously, and could be found willing to consider the importance of trustworthy research in a more personal and meaningful way. A whole-class reflective discussion around the central question '*would you dare to jump*' was staged, with follow up question such as: '*why (not)?*', and: '*could and should you have produced a scientifically more sound inquiry?*'.

In conclusion of this stage ideas were exchanged and collected on what, according to the students, constitutes a scientifically (more) sound inquiry and on the criteria that make a conclusion valuable to the stunt coordinator.

Based on the specified design intentions we can now formulate the research questions:

1. *In terms of students' intent to consider the quality of their answer to the research question in inquiry, what are the contributions of an approach that uses:*
   a) *guided inquiry combined with a context-based evaluation of the research quality,*
   b) *guided inquiry, a context-based evaluation and a change of perspective from producer to consumer of the research findings.*
2. *Once students consider the quality of their answer to the research question, what aspects of this perceived quality align with scientific quality, and which aspects are missing?*

Instruments and data collection

Data were obtained during the first stage, from (i) written work and (ii) audio recordings. In stage two it involved (iii) the submitted homework assignment. During stage three, the data sources are (iv) written answers to a reflection form and (v) audio recordings of the reflective whole-class evaluation. We present instruments (i)-(v) in turn.

i *Scientific Graphic Organiser*. During the first stage students kept track of their work in a written pre-structured lab journal known as a scientific graphic organiser (SGO) (Pols, 2019; Struble, 2007). An SGO provides a schematic for reporting the essentials of an inquiry: the research question, the chosen instruments and method, theory used, data displayed in tables and graphs, a conclusion, the argumentation supporting the conclusion and a critical evaluation.

ii *Audio recordings of the first lesson*. The teacher used an audio voice recorder to record classroom talk during the entire lesson. Salient instances, mainly pertaining to students' interpretation of the task and chosen approach, were identified and transcribed to augment the written data.

iii *Homework assignment*. Each student team wrote a letter to the stunt coordinator as discussed above, to report what they had found out about the influence of the factor they investigated on the 'swing time' of a pirate. They described how that finding came about and how trustworthy or useful they thought it was. Students' work in this stage was triangulated with the data from the conclusion and evaluation section of the SGO.

iv *Reflection form*. During the third stage, the second lesson, after the whole-class discussion on 'would you dare to jump', each student team answered the following questions in writing:
   1 What would you like to change in your investigation? Why?
   2 What do you want to achieve with that change?
   3 What makes an investigation and the written report trustworthy?

(Further questions were present in the form but have not been used in this study.)

a *Audio recordings in stage 3*. Again, the teacher recorded classroom talk during the entire lesson with an audio voice recorder. Where most of stage 1 consisted of

work in small teams, this stage included a whole-class discussion that introduced the change of role from producer to consumer of knowledge. It also included conversations during whole-class and small-team reflective activities to evaluate the quality of the inquiry. The data provide information about the effects of the change of role, the students' self-evaluations and their ideas for improvement.

Data analysis

ARPI was used to describe, analyse and rate the students' approach in the first stage, based on the choices they made in designing and executing their inquiry. Relevant information for each targeted UoE was gathered from the SGO, the letter to the stunt coordinator and the audio recordings of the first lesson. Analysis of the three data sources revealed what choices students made (e.g. regarding the number of repeated measurements) and whether they consciously substantiated these choices (e.g. with a statement such as '*since the spread in measurements is small, three repeats suffice*'). The ARPI descriptors were used to assign attainment levels to the teams and score the quality of students' actions and substantiations. For instance, for UoE 8 (Table 1) we first analysed whether students collected a single measurement (level 0), or took repeated measurements (level 2). We then investigated whether students provided a substantiation of that decision (level 4). Levels 1 or 3 were assigned when students outperformed the lower level, but did not fully reach the next level, e.g. level 3 could be assigned when measurements were repeated but an incomplete or mediocre substantiation was provided.

Application of ARPI occasionally requires a judgement call. E.g. one team first took a single measurement at a given value of the independent variable (level 0) but repeated three times subsequently (level 2). The change resulted in more reliable results in later measurements. This may reflect consideration of the quality of evidence, perhaps reflecting attainment level 4. However, they did not augment their first measurement or substantiate either their initial or later approach. In these cases we decided to err on the side of caution. In this case level 1 was assigned.

As assigning scores thus relies on an interpretation of information that is often fragmented or incomplete (students tend to be brief in specifying what is done and why), assigning students' UoE levels was carried out twice by the first author. In the few cases of mismatching scores, evidence was re-examined before a definite score for these UoE were assigned. Assigning scores was repeated by an independent, informed teacher-researcher for an arbitrarily chosen section of 30% of the dataset. The inter-rater reliability was 89%, implying that no relevant differences were found. Mismatching scores were discussed until agreement was reached.

This analysis provided an overview of the students' approach in the first stage and revealed the weaknesses in its quality from a scientific point of view. A number of UoE was not assessed as the given task did not involve their application and no relevant data could be collected (UoE 1, 3, 10, 11 17-19). As a case in point students were not required to engage in peer review, so that UoE 19 is not considered here.

To study whether the switch in perspective changed the students' perception of the usefulness and trustworthiness of their inquiry, we analysed first the level of confidence students had in their results, as expressed in the letter to the stunt coordinator. We

allocated *low (-)*, *intermediate (0)* and *high (+)* levels (or *(?)* if not expressed). Subsequently we analysed the audio recordings of the stage 3 with a focus on students' reactions and arguments when asked whether they would dare to jump. We compared their views in the letter with these verbal reactions and arguments.

Finally, students' propositions for improvement in the reflection forms were linked to UoE (RQ2). We explored the match between weaknesses they identified and those derived from a scientific perspective to determine to what extent their modified goals and intentions for change aligned with it.

All interventions, instruments and collected data were in Dutch and where necessary have been translated by the authors.

Results

First, the analysis of the scientific quality of inquiries is presented (RQ2). Next, the students' own views of that quality during the first stage of the intervention, where they plan and collect data and write to the stunt coordinator (RQ1a) are presented. Subsequently, the altered perspectives in the second stage (RQ1b) are given. Finally, data are presented on how students think their inquiry can be improved (RQ2) and compared with what is required in view of the observed scientific quality.

Students' inquiry from a scientific perspective

For each of the twelve UoE of ARPI, attainment levels of each student team were assigned on the basis of their SGO inquiry reports and letters to the stunt coordinator. The results are shown in Table 3. The student teams' operationalisation of their inquiry is analysed as follows:

ARPI phase: posing questions

UoE 2: Most teams posed a research question of the form 'find out how X influences Y', revealing that they understood what they intended to investigate. Intermediate level was assigned in cases where a relationship was not made explicit, e.g.: '*At what angle should the stuntman jump to reach the other side?*' (team G1).

ARPI phase: design

UoE 4: A relation between the experiment and the research question was often not specified. While most teams chose generally suitable instruments and procedures for measuring relevant quantities, a systematic, structured approach tended to be absent. The most extensive description was given by team G9: 'in order to see how mass influences the swing time, seven different weights (20–100 g) were used'. Another more extensive description, in the letter of G1, is presented in Figure 3.

UoE 5: Teams mostly identified variables that could potentially influence the 'swing time' and understood that therefore, these needed to be controlled (i.e. kept constant). Several failed to adequately operationalise this understanding, *e.g.* various teams increased the weight of the pendulum by hanging additional weights below one another. The ensuing discussion with the teacher showed that they understood 'fair

Table 3. Number of teams ($N = 11$) per competence level for each UoE on a 5-point scale from lowest (0) to highest (4), on average in SGO and letter. Class average level in grey. Number of teams whose UoE could not be determined in final column.

UoE			Level of competence					No score
Phase	No.	The researcher understands that	0	1	2	3	4	
Research question	2	The inquiry is an attempt to establish the relationship (or lack of one) between an independent variable and a dependent variable.	0	0	5	1	5	0
Design	4	The research question should be answerable with the devised experiment.	6	0	2	2	0	1
	5	Other variables can affect the dependent one, therefore a fair test is needed, keeping these variables constant.	2	1	5	0	3	0
	6	It is important to choose suitable instruments and procedures to get valid data with the required accuracy and precision.	10	1	0	0	0	0
	7	(Human) Errors and uncertainties may occur and precautions are needed to minimise or avoid them, ensuring reliability.	3	6	2	0	0	0
Method & procedure	8	Measured values will show inherent variation and the reliability of data must be optimised, requiring repeated measurements.	2	1	8	0	0	0
	9	The range of values of the independent variable must be wide enough and the interval small enough to ensure that a potential pattern is detectable.	3	0	3	3	1	1
Analysis	12	Data require appropriate methods for analysing and describing them.	0	2	3	2	3	1
	13	An optimally informative answer to the research question requires a description of relationships in as much detail as possible. Quantitative descriptions are more detailed than qualitative ones.	3	1	3	2	0	2
Conclusion & evaluation	14	A complete, clear, substantiated and useful answer to the research question must be formulated.	3	3	0	4	0	1
	15	The reliability of the dataset is to be accounted for by considering how well each datum was measured and the reliability of the established relationship.	1	3	5	0	0	2
	16	The validity of conclusions does not go beyond the data available. Therefore limitations to the validity of the claim should be expressed.	2	4	4	1	0	0

testing' (change only one variable at a time to establish its effect) but failed to notice that their way of increasing the weight also increased the pendulum's length.

UoE 6: This understanding is rated as 'low' for ten out of eleven teams. Teams used readily available instruments such as rulers and handheld stopwatches but did not consider the use of more accurate instruments (such as the record function on their phones) or procedures (such as measuring several swings at once instead of only a half swing at a time).

ARPI phase: method & procedure

UoE 7: The teams generally did not consider human or other measurement errors (e.g. reaction time) or procedures to address these. E.g. most failed to notice or address that the duration of the measurement they chose to do, timing half a swing, was often of the same order of magnitude as the measurement error caused by their reaction time.

UoE 8: Teams tended to repeat measurements a fixed number of times (usually 3) but without any suggestion of an understanding that this would suffice to take inherent variation into account and thus enhance the findings' reliability. Since their action were most likely routine rather than reasoned, an intermediate competence level was assigned.

UoE 9: Three teams chose an inadequate range or interval for their measurements, e.g. using a range of a few centimetres within the available range of the meter-long pendulum.

ARPI phase: analysis

UoE 12: As is shown for example in Figure 2, most students created data representations that allowed for the identification of a pattern (if present).

UoE 13: They were unable to describe the pattern in the data, if one was found, quantitatively. Minute differences in measured values were regularly seen as significant.

ARPI phase: conclusion & evaluation

UoE 14: In line with the quality of the dataset and its analysis, the conclusions and evaluations were brief and superficial. Some illustrative examples in SGO's and letters are:

G3: The lighter the weight, the shorter the swing time, so it seems.
G6: The difference per rope (material) is minimal, but of importance for timing the perfect jump.
G10: The bigger the (starting) angle, the longer the swing time, but noticeably only from 40° onwards. It doesn't differ much, but it is clear.

These qualitative conclusions did not meet scientific requirements, they were insufficiently informative and not useful from the perspective of the given context. Especially in cases where the relation is not present or measurable (rope material, mass) or not evident

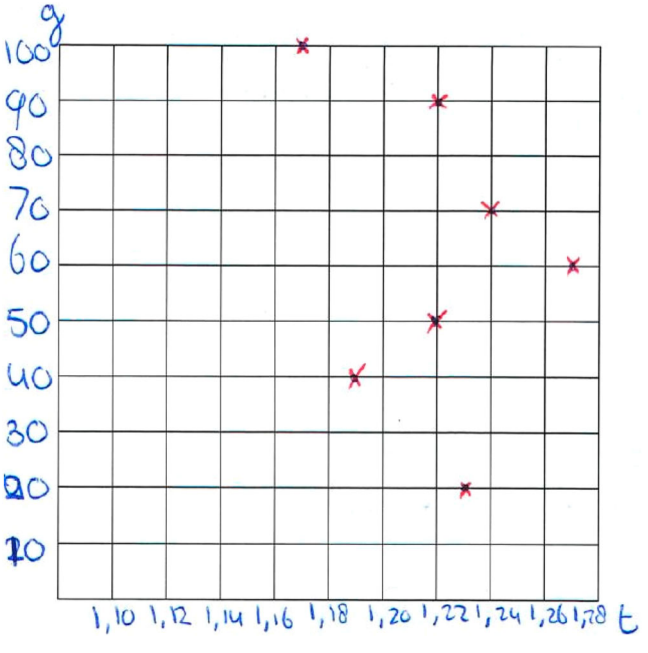

Figure 2. Students interpreted their well-presented data as showing that an increase in mass results in a larger period, although the variation in the measured period is within the margin of error and this inference unwarranted.

(angle), students had difficulties in describing the effect of the variable under investigation.

Students' initial perspective of their inquiry in stage 1

Recorded exchanges between the teacher and some teams suggest that they took the contextualisation of the pendulum in terms of the pirates' swing seriously, and viewed a high-quality answer as required in this investigation. While students were executing their plan and carrying out measurements, the teacher asked teams whether the stuntman could have confidence in their work. Two of these illustrative exchanges follow.

Exchange with G9:

Teacher:	How is it going?
Lisa:	I think it is going fine, but I am not sure. We are using a weight of 20 g, we will do the same measurement for 50 g. That is correct, right?
Teacher:	Yes, seems reasonable. Would the stuntman have confidence in the research?
Lisa:	Yes.
Teacher:	Why?
Jolien:	It is scientific.
Lisa:	We try to do it as scientifically as possible.

Exchange with G5:

Teacher:	Do you think a stuntman would have confidence in your research?
Masha:	I think so.
Teacher:	Why?
Masha:	Because all the measurements are more or less the same, so the measurements were fine.

Most other teams provided less clear and concrete answers: '*we are still figuring things out*'. However, G9 and G5 are evidently confident about their plan, either because they believe that they are using a scientific approach, or observe minimal variability in their measurements.

In line with the findings in Table 3, most carried out the inquiry as is often reported in the literature: quickly and without explicit consideration of the quality of data. While some students genuinely believed they tried as hard as they could, they did not feel the urge to ask the teacher for better instruments or methods to determine the swing time.

Students' perspective of inquiry in the letter to the stunt coordinator in stage 2

Instead of ending the inquiry with the writing of a report, stage two of the intervention was initiated. Students were to justify their work and rate their confidence in their findings by writing a letter to the stunt coordinator, in response to the fictional request for help. Their level of confidence is interpreted as indicative of their perception of the quality of their inquiry. Did this context cause the students to adjust their perspective of inquiry and of the quality of their findings? Two examples of complete letters are shown in Figure 3, the other letters are included in the journal's data repository.

Four of the eleven teams (G1,G2,G4,G9) stated that they were not confident that the inquiry findings could be used, four teams had some confidence (G3,G7,G10,G11) and

two teams (G5,G8) explicitly stated they did have confidence in their own findings. Team G6 did not mention its level of confidence. Notably, their lack of confidence was not due to a lack of effort. The teams did feel they tried to produce quality research (as exemplified in the underlined sections below) but that they encountered 'insuperable', externally attributed problems:

G6: We measured 4x per rope to increase the accuracy of the measurement. It is hard to measure accurately, but *we tried the best we could*. (…) We *hope to have helped* you.
G9: We *tried our utmost*, but because we did not have equipment to measure very precisely, we are, unfortunately, not confident that our research findings are useful.

Note that the two letters presented in Figure 3 also present a (justifiable) lack of confidence.

Students' view of inquiry in the reflective discourse in stage 3

In the third stage of the study, after submitting their letters, students were asked to change their perceived role from researcher to stunt(wo)man, which was meant to provide them with a new perspective and reconsideration of the quality of their findings. This is how students responded to the teacher's introduction in stage 3:

Teacher: Suppose you are the stuntman standing on the edge of the ship, you have a 12 m long, 5 cm thick rope in your hands and you have to jump soon. Just before the stunt, you have read how the stunt should be performed. The

G1 We have investigated how the starting angle of the sling with the stuntman influences the swing time. To that purpose we have used a rope and a small weight attached to it. We started with an angle of 10 degrees, recorded the time from release to dead centre using a stopwatch. We repeated this with angles of 20, 30, 40, 50, 60 and 70 degrees.

We could not find major differences in swing time, as the period was 0.60 s at an angle of 10 degrees and 0.58 at an angle of 40 degrees.

This leads us to conclude that the starting angle hardly influences the swing time. If you start at an angle of 10 degrees, your velocity will be higher as you start higher, but you will travel more distance as well. At an angle of 70 degrees, your velocity will be lower, but you travel less far, as a result the swing time is roughly the same.

We are not yet confident that you can use our findings as our study is limited in scope. We advise to use a bigger set up with a rope of 2 or 3 meters in length.

G2 We have investigated how the length of the rope influences the swing time. We have found that the length indeed affects the swing time, but this happens because the distance travelled changes as well. The longer the rope, the more distance is covered, the longer the swing time. We are not confident that you can use our results because, firstly, it is not precise. It is also self-evident that the swing time increases when the rope is longer as the distance travels increases.

Figure 3. Two exemplary letters to the stunt coordinator in which students explain what has been done and found and whether they have confident in the quality of their own inquiries.

	stunt is based on your own reports and investigations … Would you dare to jump?
Lisa (G9):	(interrupts) No.
Teacher:	Why not?
Lisa (G9):	It is not measured with proper equipment, it is based on us … you do not know how and what, exactly.
Teacher:	You think your measurements are not adequate enough?
Lisa (G9):	No.
Teacher:	And that is due to the equipment?
Lisa (G9):	And ourselves, you cannot start from the exact same starting point each time. And the equipment is not good, better equipment is required.
Teacher:	So what do you suggest? What do you want to improve?
Lisa (G9):	Every time using the same point for your measurement.
Teacher:	The angle at which you start you mean?
Lisa (G9):	Yes, and where you start and stop timing.

As no other teams responded, the teacher asked again who would dare to jump:

Teacher:	Who would dare to jump?
Thim (G10):	Sure, why not? (other students are laughing)
Teacher:	Sure? You trust in what you have done?
Tom (G10):	If the rope is tightened. You can always swing back.
Teacher:	But, what happens if you're too early?
Bob (G8):	BOOOM.
Teacher:	Boom, you will land in the explosion. This brings a potential risk. Do you still consider that you have produced a sound study?
Thim (G10):	Our calculations are correct.
Teacher:	Who does not trust their own inquiry? (pause) Silvester?
Silvester (G7):	Yes, what Lisa says.
Teacher:	Could you have done better?
Silvester (G7):	I think so, yes. Measuring time accurately was difficult.

Thim and his partner Tom seemed to have confidence in their findings. However, in their earlier letter to the stunt coordinator they qualified these less decisively as 'reasonably reliable'. All other students agreed with Lisa and deemed the quality of the inquiry insufficient in light of the risk of being hurt.

In this class, the design intention of effecting a change in the students' evaluation of their inquiry was instantiated. Both in Lisa's concerned consternation, Thim's brazen indifference, and the verbal and non-verbal responses of the rest of the class that are harder to convey, students are seen to recognise that actually using their findings could cause harm.

Capitalising on their fresh perspective, the teacher fostered students' development of quality criteria for conclusions in inquiry. Presenting once again their earlier conclusions in order of increasing precision and detail (but without revealing that), students contemplated what characterises that quality. The teacher asked whether the conclusion '*the length of the swing affects the swing time*' helps the stunt coordinator design the stunt. Although some said yes, one student convinced the others that it is not helpful since it is not specified whether a shorter rope results in a shorter or longer swing time. Several students regarded '*the longer the rope, the longer the swing time*' as useful until the teacher asked how this conclusion would help them calculate the swing time for a 12 m long rope. Yet another possible conclusion was therefore forwarded by the teacher:

Teacher: If the rope is 4x as long, the swing time is doubled.
Lisa (G9): Yes.
Teacher: What do you mean?
Lisa (G9): That will help you.
Teacher: Why?
Lisa (G9): You have numbers. You can make a prediction based on the numbers.

While many conclusions tend to fit the data of an investigation, its purpose is to find the most useful conclusions, which is optimally specific. Developing this understanding in students was an aim of this discussion. An exchange that immediately followed suggests that it was likely to have been attained:

Teacher: What do you learn from this about drawing conclusions?
Thim (G10): You really have to think about the conclusions.
Teacher: I guess so. Why?
Tom (G10): Otherwise it is of no (expletive) use to the stunt coordinator. … He can't do anything with that.

Students' written reflection on what is learned

In order to consolidate the insights gained from the exchange, students reviewed their work in answering the open questions of the reflection forms and offered recommendations to improve their inquiry. This reflective activity was meant to foster students' metacognitive development, their insight into what they learned and how they learned it.

In describing what they would like to change in their investigation, why, and to what purpose students proposed, *e.g.* different methods of measuring the swinging time more accurately:

> G1: Use a larger longer rope, this increases the swing time and makes it therefore easier to accurately measure the time. Use a sensor to measure when the swing is released and stops when the swing is at the other side. This way you don't have to deal with reaction time and thus results in a more accurate measurement. Attach the triangle ruler to the setup in order to measure angles accurately.

> G3: We would like to use professional equipment for obtaining measurements. We probably did not measure and calculate everything perfectly resulting in findings that are not quite right. What we want to achieve with this is that we can optimize our conclusion and the stunt can be performed in a safe way.

In all instances, teams identified weaknesses in their inquiries. Their ideas and thoughts show a lack of experience in inquiry but *accord* with scientific criteria for improving the quality of their investigation. Students' replies to further reflective questions, as to what makes inquiry results trustworthy or of good scientific quality, or what they learned from doing the inquiry, tended to repeat these answers but without providing further insights, e.g.:

> G1: Many & accurate measurements (UoE7&8). Good and substantiated explanation (UoE14). Good elaboration. Professional equipment and instruments (UoE6).

> G3: If the inquiry is carried out professionally and seriously, with good, reliable equipment. You need to check whether the data are correct.

Students' answers, illustrated by these examples, showed that students' notion of a trustworthy inquiry accords with a scientific perspective. However, their ideas lack practical detail and clarity in terms of operationalisation. E.g. in 'many measurements', how many are meant? The following exchange, occurring towards the end of the lesson, illustrates what students said to have learned about doing scientific inquiry:

Teacher:	What rules have you learned? Have you learned any?
Eric (G3):	Yes. Well, you really have to think.
Teacher:	About what?
Eric (G3):	About the conclusion.
Teacher:	Anything else?
Eric (G3):	That after a single measurement you don't just have a measurement right away. That you have to measure several times before you have a good measurement.
Teacher:	Well, these are two lovely things you have learned. Why do you want several measurements?
Eric (G3):	Well, if you take a measurement, and that measurement is not good, then you have a wrong measurement and then the stunt can go wrong.

While students were not yet able to specify in detail what they had learned, their words in our view implied the understanding that inquiry is meant to render not just an answer to the research question, but the best possible answer in the given circumstances. They expressed 'the best possible answer' in terms of trustworthiness and usefulness. They provided reasons and examples from the inquiry's context to explain why such an answer required.

Discussion

Using the ideas of Millar et al. (1994), we assume that students may start to make scientifically desirable choices in inquiry independently once they understand that inquiry needs to aim at producing the best possible answer given the circumstances. Therefore we tried in this study to have them consider *the value of scientific quality* of their inquiry first, before further developing understanding of how to produce that quality. We discuss below whether and when we succeeded, and the extent to which design intentions were attained.

Answers to the research question

Stage 1 of the practical involved the deceptively simple physical pendulum (Matthews, 2001) but deviated from the conventional 'cookbook' exercise to confirm the formula relating length to period. As suggested by various scholars, we gave students more agency of their inquiries (Crawford, 2014, p. 527; Hofstein & Kind, 2012; Zion & Mendelovici, 2012). We encouraged them to forward their own ideas about factors that might influence the period, and to study these as they saw fit. Reducing the cognitive load in terms of knowledge of types **A**, **B** and **C** of the PACKS model allowed students to focus on the aspects involved in knowledge type **D**: their use of criteria involved in the construction and evaluation of scientific evidence. We observed, however, that context-based, guided inquiry and explicit self-evaluation of the research quality did

not sufficiently affect the students' intent (RQ1a). For example, when asked to consider the quality of their answer to the research question (Q: '*Would the stuntman have confidence in the research?*'), students understood that quality to be adequate in a scientific sense (A: '*Yes, because it is scientific*'). The students' words and actions did not sufficiently reflect the understandings that are required to render scientific adequacy to evidence in inquiry (Table 3). For example, they chose the first inquiry methods and approaches that came to mind without searching or asking for better alternatives. 'Better', that is, in terms of criteria they themselves formulated later on in stage 3, but not during stage 1. With very few exceptions, the guided and contextualised character of the activity does not sufficiently foster students' awareness of the value of a scientific approach (RQ 1a), confirming findings of e.g. Molyneux-Hodgson et al. (1999).

Stage 2 emphasised the context again as students were asked to write a letter to the stunt coordinator. The letters showed that students were either still quite content with the quality and nature of their conclusions or that they, partially, deflected responsibility for the quality of the findings ('*We did not have equipment to measure very precisely*.'). Their perspective on the inquiry was that of a 'scientific investigation in a classroom context' (Millar et al., 1994), i.e. with the purpose of finding an answer to the research question but no personalised criteria for the scientific quality of that answer.

In stage 3 of the practical, in answer to RQ1b, we explored whether a change in the students' perspective from producer to consumer of the research findings can foster their (re)consideration of the quality of the inquiry. In considering the *practicality of consequences* of their findings in a new way students came to the view that developing 'trustworthiness' and 'usefulness' *ought* to be demanded of the answer to the research question but were – according to their own standards – not yet achieved. As students acknowledged that the inquiry should have been performed differently, they explored in a guided way what should be changed. The teacher selected and presented the conclusions of the different teams, in order of increasing precision and detail. Students were able, collectively, to identify these ordering criteria and to interpret them as making the answer more useful and trustworthy, therefore preferable.

From students' own ideas about how the quality of their inquiry could be improved, we can infer which aspects of this perceived quality align with scientific quality, and which aspects are missing (RQ2). We conclude that students' own suggestions for improvements, derived from their reflection forms, all aligned with and could be interpreted in terms of the UoE of Table 3. In a qualitative, general sense this signifies that a cognitive motive was now present for developing UoE related to the adequate collection and analysis of data, and formulating an adequate conclusion. As was expected, they were unable to provide sufficient detail and clarity to operationalise their ideas. They seemed to see the point of adhering to several of the UoE in inquiry, because doing so contributes to the trustworthiness and usefulness of the findings. However, as was expected, they were not quite able to explain the underlying scientific standards, or the methods used to satisfy these.

Implications

According to the literature, students in inquiry (seem to) act almost without thinking, (seemingly) indifferent to establishing a valid and reliable answer to the research

question, or ignorant of how to obtain it. This study shows, however, that even if students appear interested, motivated and engaged: they fail to see the point of obtaining better answers and lack criteria for evaluation of the quality of such answers. in In making practical work more effective and enabling students to engage in basic scientific inquiry (Abrahams et al., 2013; Hodson, 2014; Hofstein & Kind, 2012) we direct students' attention to the value and purpose of scientific investigations. The question of why some answers are better than others, and what is meant by 'better' in science, appears to be a useful starting point for learning the methods and techniques scientists apply to optimise the quality of their inquiries. As shown, appealing to students' empathy and encouraging them to develop personally relevant criteria is one way to do so. The combination of context, reflection and a change of perspective from *producer to consumer of knowledge* contributes to an educational design that accomplishes this. While the intuitive concepts 'trustworthiness' and 'usefulness' are not necessarily fully developed in a scientific sense, they align with and can be developed further into the more fundamental but abstract concepts of reliability and validity.

This study has implications for integrating argumentation into inquiry, advocated by influential authors (Erduran & Jiménez-Aleixandre, 2008; Gott & Duggan, 2007; Newton et al., 1999; Osborne, 2013) but scarce in terms of empirical studies attempting it (Driver et al., 2000; Erduran & Jiménez-Aleixandre, 2008; Watson et al., 2004). We have argued elsewhere that conducting inquiry can be interpreted as the construction of an optimally cogent argument in support of an optimally informative claim on the basis of optimally valid and reliable data (Pols et al., 2022). Engaging in argumentation requires students to have a notion of what counts as scientifically cogent, i.e. of what makes some answers to research questions better, in a scientific sense, than others. This study provides an example of a starting point for developing these notions and satisfying the preconditions for students engaging in argumentation. We have provided an example of students' successful argumentation in establishing the most informative answer to their research question.

Limitations and further research

More research is needed to explore how the learning effects in this intervention can be consolidated and utilised in the further developments described above. As a first step, the collection of UoE in Table 3 has been validated as a set of norms and standards by which the quality of virtually all students' inquiry in physics can be assessed. This set of UoE is suitable in guiding student-researchers in developing or evaluating that quality, and in argumentation aimed at the construction or evaluation of the scientific cogency of a researcher's claims. Developed and validated with physics students at BSc level, the next step will be to develop learning pathways for levels between that of the current study and university level. As a starting point, a teaching sequence was developed targeting a range of the UoE that integrates the current intervention. It explores the further development of inexperienced students' intuitive concepts in inquiry learning and argumentation.

Further research is needed to establish whether the findings obtained in this small-scale, qualitative and exploratory study can be replicated at a larger scale, and explore conditions that render ecological validity to the design. For example, a crucial yet vulnerable element of the activity is the acceptance of the realistic but entirely fictitious context.

We did not investigate what conditions are sufficient or necessary to create a classroom environment where this acceptance of role play can occur. Obviously, the teacher plays an important role in fostering the essential mutual respect and trust but further conditions may have to be satisfied to prevent students from dismissing the role play as childish or 'fake'. As it is known that many teachers are not well equipped to give substance to the learning goal *learning to engage in scientific inquiry* (Abrahams et al., 2014; Abrahams & Millar, 2008; Crawford, 2014; Lunetta et al., 2007; Smits, 2003), a question remains whether similar results can be obtained by other teachers. Anecdotal data are available in this respect from four teachers in our network who were inspired by the activity and tried it out in their own classes. In three of their informal reflective reports, we found the observed learning to align largely with what is reported here, while in one case students refuted the context and did not acquire the intended understandings. Creating conditions where role play in teaching is taken seriously and rendered effective is a topic for further research.

Conclusion

Recently, Hofstein (2017); Najami et al. (2020) stated again that 'the biggest challenge for practical work, historically and today, is to change the practice of "manipulating equipment not ideas"'. We investigated whether having students repeatedly consider the context of the inquiry instigates them to evaluate and improve the quality of their approach, turning the hands-on into a minds-on activity. We established that students may enjoy and work hard in contextualised inquiry that involves explicit self-evaluation of the quality of their work. However, this in itself does not enable them to adopt a critical view on the quality of their approach. Students accepted the purpose of inquiry as 'finding an answer to the research question', but, in accord with the literature and our professional experience, seemed happy with any answer they could find.

They did adopt that more critical view when asked to change their perspective from that of the researcher producing knowledge to that of the consumer of that knowledge, considering hypothetical exposure to the potentially harmful implications of utilising that knowledge. Their personal purpose of inquiry changed from 'finding any answer to the research question' to 'finding the most trustworthy and useful answer obtainable with the means and the time available to us'. Future research will be directed at exploring ways to develop this notion further, towards 'finding – the most informative, reliable and valid answer to the research question within the given constraints and limits imposed by feasibility of obtaining it' and develop the procedural and conceptual knowledge that enables them to find that answer (Pols et al., 2022). We intend to explore how to further develop this mental readiness, the personal cognitive needs and the inquiry knowledge in a learning process aimed at obtaining answers of this kind. We think it may foster an eagerness in students to apply scientific standards in inquiry without having to be told to do so.

Materials

The educational materials are available in English, Dutch, French, Spanish and Basque (Pols, 2021a). For more details on the teaching sequence, see Pols et al. (2019).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is part of a research programme for teachers financed by the Netherlands Organisation for Scientific Research (NWO) (Grant Number 023.003.004).

Data availability statement

The letters to the stunt coordinator and students' written response to the reflective questions are translated and available through the journal's data repository.

ORCID

C.F.J. Pols http://orcid.org/0000-0002-4690-6460

References

Abrahams, I. (2005). Between rhetoric and reality: The Use and effectiveness of practical work in secondary school science [Unpublished doctoral dissertation]. UK: University of York.

Abrahams, I. (2011). *Practical work in secondary science: A minds-on approach*. Continuum.

Abrahams, I., & Millar, R. (2008). Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education*, *30*(14), 1945–1969. https://doi.org/10.1080/09500690701749305

Abrahams, I., Reiss, M. J., & Sharpe, R. (2013). *Improving the assessment of practical work in school science: lessons from an international comparison*. York: https://www.gatsby.org.uk/uploads/education/reports/pdf/improving-the-assessment-of-practical-work-in-school-science.pdf

Abrahams, I., Reiss, M. J., & Sharpe, R. (2014). The impact of the 'Getting Practical: Improving practical work in science'continuing professional development programme on teachers' ideas and practice in science practical work. *Research in Science & Technological Education*, *32*(3), 263–280. https://doi.org/10.1080/02635143.2014.931841

Bakx, A., Bakker, A., Koopman, M., & Beijaard, D. (2016). Boundary crossing by science teacher researchers in a PhD program. *Teaching and Teacher Education*, *60*, 76–87. https://doi.org/10.1016/j.tate.2016.08.003

Banchi, H., & Bell, R. (2008). The many levels of inquiry. *Science and Children*, *46*(2), 26. https://www.michiganseagrant.org/lessons/wp-content/uploads/sites/3/2019/04/The-Many-Levels-of-Inquiry-NSTA-article.pdf

Bennett, J., Lubben, F., & Hogarth, S. (2007). Bringing science to life: A synthesis of the research evidence on the effects of context-based and STS approaches to science teaching. *Science Education*, *91*(3), 347–370. https://doi.org/10.1002/sce.20186

Brewer, M. B. (2000). Research design and issues of validity. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16). Cambridge University Press.

Bruckheimer, J. (2007). Pirates of the Caribbean: at world's end. *Pirates of the Caribbean*. https://youtu.be/SfyePrFKvVA

Crawford, B. A. (2014). From inquiry to scientific practices in the science classroom. In N. G. Lederman, & S. K. Abell (Eds.), *Handbook of research on science education (Vol. 2)* (pp. 515–541). Routledge.

de Putter-Smits, L. G. A. (2012). *Science teachers designing context-based curriculum materials: developing context-based teaching competence* [Unpublished doctoral dissertation]. https://doi.org/10.6100/IR724553.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312. https://doi.org/10.1002/(SICI)1098-237X(200005)84

Duschl, R. (2000). Making the nature of science explicit. In R. Millar, J. Leach, & J. Osborne (Eds.), *Improving science education: The contribution of research* (pp. 187–206). Open University Press.

Erduran, S., & Jiménez-Aleixandre, M. P. (2008). *Argumentation in science education. Perspectives from classroom-based research*. Springer.

Glaesser, J., Gott, R., Roberts, R., & Cooper, B. (2009). Underlying success in open-ended investigations in science: Using qualitative comparative analysis to identify necessary and sufficient conditions. *Research in Science & Technological Education*, *27*(1), 5–30. https://doi.org/10.1080/02635140802658784

Gott, R., & Duggan, S. (1996). Practical work: Its role in the understanding of evidence in science. *International Journal of Science Education*, *18*(7), 791–806. https://doi.org/10.1080/0950069960180705

Gott, R., & Duggan, S. (2003). *Understanding and using scientific evidence: How to critically evaluate data*. Sage Publications Ltd.

Gott, R., & Duggan, S. (2007). A framework for practical work in science and scientific literacy through argumentation. *Research in Science & Technological Education*, *25*(3), 271–291. https://doi.org/10.1080/02635140701535000

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (2003). *Research into understanding scientific evidence*. http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm

Hodson, D. (1990). A critical look at practical work in school science. *School Science Review*, *70*(256), 33–40. https://eric.ed.gov/?id = EJ413966

Hodson, D. (1992). Assessment of practical work. *Science & Education*, *1*(2), 115–144. https://doi.org/10.1007/BF00572835

Hodson, D. (2014). Learning science, learning about science, doing science: Different goals demand different learning methods. *International Journal of Science Education*, *36*(15), 2534–2553. https://doi.org/10.1080/09500693.2014.899722

Hofstein, A. (2017). The role of laboratory in science teaching and learning. In K. S. Taber, & B. Akpan (Eds.), *Science education* (pp. 357–368). Springer.

Hofstein, A., & Kind, P. M. (2012). Learning in and from science laboratories. In B. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 189–207). Springer.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*(1), 28–54. https://doi.org/10.1002/sce.10106

Holmes, N. G., & Wieman, C. (2016). Examining and contrasting the cognitive activities engaged in undergraduate research experiences and lab courses. *Physical Review Physics Education Research*, *12*(2), 1–11. https://doi.org/10.1103/PhysRevPhysEducRes.12.020103.

Holmes, N. G., & Wieman, C. (2018). Introductory physics labs: WE CAN DO BETTER. *Physics Today*, *71*(1), 1–38. https://doi.org/10.1063/PT.3.3816.

Johnstone, A. H., & Wham, A. (1982). The demands of practical work. *Education in Chemistry*, *19*(3), 71–73.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769. https://doi.org/10.1002/tea.20020

Kortland, J. (2007). *Context-based science curricula: Exploring the didactical friction between context and science content*. Paper presented at the ESERA 2007 Conference, Malmö, Sweden. www.phys.uu.nl/~kortland>English>Publications

Lijnse, P. (2014). *Omzien in verwarring*. Fisme.

Lipton, P. (2003). *Inference to the best explanation*. Routledge.

Lunetta, V. N., Hofstein, A., & Clough, M. P. (2007). Learning and teaching in the school science laboratory: An analysis of research, theory, and practice. In N. Lederman, & S. K. Abell (Eds.), *Handbook of research on science education* (pp. 393–441). Lawrence Erlbaum Associates.

Matthews, M. R. (2001). How pendulum studies can promote knowledge of the nature of science. *Journal of Science Education and Technology*, *10*(4), 359–368. https://doi.org/10.1023/A:1012299219996

Millar, R. (2004). *The role of practical work in the teaching and learning of science*. National Academy of Sciences.

Millar, R., Le Maréchal, J. F., & Tiberghien, A. (1999). Mapping the domain: Varieties of practical work. In J. Leach, & A. Paulsen (Eds.), *Practical work in science education - Recent research studies* (pp. 33–59). Roskilde University Press/Kluwer.

Millar, R., Lubben, F., Gott, R., & Duggan, S. (1994). Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance. *Research Papers in Education*, *9*(2), 207–248. https://doi.org/10.1080/0267152940090205

Molyneux-Hodgson, S., Sutherland, R., & Butterfield, A. (1999). Is 'Authentic'Appropriate? The Use of work contexts in science practical activity. In J. Leach, & A. Paulsen (Eds.), *Practical work in science education: Recent research studies* (pp. 160–174). Kluwer.

Najami, N., Hugerat, M., Kabya, F., & Hofstein, A. (2020). The laboratory as a vehicle for enhancing argumentation among pre-service science teachers. *Science & Education*, *29*(2), 1–17. https://doi.org/10.1007/s11191-020-00107-9

Netherlands Institute for Curriculum Development. (2016). http://international.slo.nl

Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education*, *21*(5), 553–576. https://doi.org/10.1080/095006999290570

Ntombela, G. (1999). A marriage of inconvenience? School science practical work and the nature of science. In J. Leach, & A. C. Paulsen (Eds.), *Practical Work in Science Education: Recent Research Studies* (pp. 118–133). Netherlands: Springer.

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

Ottevanger, W., Oorschot, F., Spek, F., Boerwinkel, D.-J., Eijkelhof, H., de Vries, M. J., … Kuiper, W. (2014). *Kennisbasis natuurwetenschappen en technologie voor de onderbouw vo: Een richtinggevend leerplankader*: SLO (nationaal expertisecentrum leerplanontwikkeling).

Pols, C. F. J. (2021b). What's inside the pink box? A nature of science activity for teachers and students. *Physics Education*, *56*(4), 045004-1–045004-6. https://doi.org/10.1088/1361-6552/abf208.

Pols, C. F. J., Dekkers, P. J. J. M., & de Vries, M. J. (2019). Introducing argumentation in inquiry—a combination of five exemplary activities. *Physics Education*, *54*(5), 410–411. https://doi.org/10.1088/1361-6552/ab2ae5.

Pols, C F J, Dekkers, P J J M, & De Vries, M J. (2022). Defining and Assessing Understandings of Evidence with Assessment Rubric for Physics Inquiry - Towards Integration of Argumentation and Inquiry. *Phys. Rev. Phys. Educ. Res*, *18*(1). https://doi.org/10.1103/PhysRevPhysEducRes.18.010111

Pols, C. F. J., Dekkers, P. J. J. M., & de Vries, M. J. (2022). Defining and assessing understandings of evidence with assessment rubric for physics inquiry - towards integration of argumentation and inquiry. *Physical Review Physics Education Research*, *18*(010111-1), 1–010111-7. https://doi.org/10.1103/PhysRevPhysEducRes.18.010111.

Pols, C. F. J. (2019). De scientific graphic organizer. *NVOX*, *44*(8), 410–411.

Pols, C. F. J. (2021a). *A teaching sequence on physics inquiry*. https://zenodo.org/record/5761998#.Ya41c7rTVPY

Pols, C. F. J., Dekkers, P. J. J. M., & de Vries, M. J. (2021). What do they know? Investigating students' ability to analyse experimental data in secondary physics education. *International Journal of Science Education*, *43*(2), 1–24. https://doi.org/10.1080/09500693.2020.1865588

Smits, T. J. M. (2003). *Werken aan kwaliteitsverbetering van leerlingonderzoek: Een studie naar de ontwikkeling en het resultaat van een scholing voor docenten*. CD-ß Press, Centrum voor Didactiek van Wiskunde en.

Spek, W., & Rodenboog, M. (2011). *Natuurwetenschappelijke vaardigheden onderbouw havo-vwo*: SLO, nationaal expertisecentrum leerplanontwikkeling.

Struble, J. J. S. S. (2007). Using graphic organizers as formative assessment. *Science Scope*, *30*(5), 69–71. https://www.nsta.org/science-sampler-using-graphic-organizers-formative-assessment

Tamir, P. (1991). Practical work in school science: An analysis of current practice. In B. E. Woolnough (Ed.), *Practical science* (pp. 13–20). Open University press.

van den Berg, E. (2013). The PCK of laboratory teaching: Turning manipulation of equipment into manipulation of ideas. *Scientia in Educatione*, *4*(2), 74–92. https://ojs.cuni.cz/scied/article/download/86/72/0

Watson, R., Swain, J. R., & McRobbie, C. (2004). Students' discussions in practical scientific inquiries. *International Journal of Science Education*, *26*(1), 25–45. https://doi.org/10.1080/0950069032000072764

Wieman, C. (2015). Comparative cognitive task analyses of experimental science and instructional laboratory courses. *The Physics Teacher*, *53*(6), 349–351. https://doi.org/10.1119/1.4928349

Zion, M., & Mendelovici, R. (2012). Moving from structured to open inquiry: Challenges and limits. *Science Education International*, *23*(4), 383–399. http://files.eric.ed.gov/fulltext/EJ1001631.pdf