

Integrating argumentation in physics inquiry: A design and evaluation study

C. F. J. Pols[✉],* P. J. J. M. Dekkers[✉], and M. J. de Vries

*Delft University of Technology, Department of Science Education and Communication,
Lorentzweg 1, 2628 CJ Delft, Netherlands*

 (Received 22 June 2023; accepted 20 November 2023; published 28 December 2023)

This small-scale, qualitative study uses educational design research to explore how focusing on argumentation may contribute to students' learning to engage in inquiry independently. Understanding inquiry as the construction of a scientifically cogent argument in support of a claim may encourage students to develop personal reasons for adhering to scientific criteria and to use these with understanding rather than by rote. An understanding of the characteristics of scientific evidence may clarify *why* doing inquiry in specific ways is important, in addition to the *how*. On the basis of five design principles—derived from literature—that integrate argumentation in inquiry and enhance learning through practical activities, we developed a teaching-learning sequence of five activities aimed at developing inquiry knowledge in lower secondary school students. By means of observations of a grade 9 physics class ($N = 23$, aged 14–15), students' answers to worksheets, and self-reflection questions, we explored whether the design principles resulted in the intended students' actions and attitudes. We studied whether the activities stimulated students to engage in argumentation and to develop the targeted inquiry knowledge. The focus on argumentation, specifically through critical evaluation of the quality of evidence, persuaded students to evaluate whether what they thought, said, or claimed was “scientifically” justifiable and convincing. They gradually uncovered key characteristics of scientific evidence, understandings of what counts as convincing in science, and why. Rather than adopting and practicing the traditional inquiry skills, students in these activities developed a cognitive need and readiness for learning such skills. Of their own accord, they used their gained insights to make deliberate decisions about collecting reliable and valid data and substantiating the reliability of their claims. This study contributes to our understanding of how to enable students to successfully engage in inquiry by extending the theoretical framework for argumentation toward teaching inquiry and by developing a tested educational approach derived from it.

DOI: [10.1103/PhysRevPhysEducRes.19.020170](https://doi.org/10.1103/PhysRevPhysEducRes.19.020170)

I. INTRODUCTION

Enabling students to engage in independent scientific inquiry is a highly valued but seemingly elusive goal of science education [1–9]. Secondary school physics students are typically meant to acquire the associated competences through engaging in quantitative physics inquiry¹: In small teams of 2–4, they manipulate instruments and materials to answer a given research question [10] which often is of the form “What is the mathematical relationship between X and Y?” (see, e.g., [11–14]). The extensive literature reports that students generally fail to use the rules and

procedures for obtaining optimally reliable and valid data in inquiry unless explicitly instructed on what to do [10,15,16]. Even motivated, interested, and able students rarely independently make methodological decisions adequately [10,17–19]. They hardly attend to what they do and why [9,20] or consider how to improve the quality of the outcomes [9,21–23].

In this study, we propose that the root cause for these problems is that students interpret their task as: *find an answer to the research question, any answer will do* [15]. This interpretation often leads to students knowing what to do and knowing how to do it, but still failing to do so (e.g., taking sufficient repeated readings). As Osborne [24] noted, “it is not just a matter of knowing how to get reliable data, but also why reliability and validity are important.”

If the above is correct, then—we think—a major step in enabling students to engage in independent scientific inquiry can be made by changing their perception of the purpose of physics inquiry first. They *ought* to interpret their task as real physicists: *find the best possible answer to the research question within the given practical constraints* [15,25,26]. Next, we ought to teach

*Corresponding author: c.f.j.pols@tudelft.nl

¹We focus in the article only on this type of inquiry. The terms quantitative physics inquiry, physics inquiry, and inquiry are used interchangeably.

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

them what qualifies as “best possible” in physics. When they later engage in inquiry, their aim becomes to convince themselves that they have reached the best possible answer. They then subject their inquiry to scrutiny by others (e.g., the teachers), trying to convince them that their claim is valid [27].

Since “becoming convinced” involves the evaluation of arguments, we see an integration of inquiry and argumentation as a potential way forward in addressing the problems in enabling students to engage in inquiry. To devise and conduct a physics inquiry well, students need—in addition to a thorough understanding of the purpose of inquiry—the ability to use argumentation adequately to guide their inquiry toward this best possible answer [15,25,26,28] and sufficient reason to obtain it. While advocated variously in the literature [27,29–32], the integration of inquiry and argumentation (to our knowledge) has not been subjected to empirical study. This leads us to the central research question of this paper:

What does integrating argumentation in teaching inquiry contribute to student understanding, critical attitude, and use of argumentation in doing quantitative physics inquiry?

A. This study

In this study, we developed and tested a teaching-learning sequence of five activities aimed at developing inquiry knowledge in lower secondary school students. Specific learning goals were derived from a previous study on the integration of argumentation and scientific inquiry [25]. These goals, encompassing basic but fundamental insights, are considered essential for enabling young students to undertake independent inquiry. To align with these identified learning goals, we carefully selected activities from both existing literature [33–35] and our own instructional practices [36]. Employing design principles—derived from literature—that integrate argumentation in inquiry and enhance learning through practical activities, we then redesigned these activities. This yielded a teaching sequence in which we first have students consider that the quality of their data—forming the basis of scientific evidence—is crucially relevant. Then we focus their attention on several of the common understandings used to gauge that quality—understandings that scientific arguments in support of research claims are based upon. Finally, we encourage them to apply these understandings in their own inquiry, to guide their choices in constructing and justifying optimally cogent answers to their research questions. This study is meant to establish the effectiveness of guidelines for the design of activities that integrate argumentation with inquiry in this way.

II. THEORETICAL FRAMEWORK

We discuss the central role of argumentation in scientific inquiry using the structure of an argument of the Toulmin

model [37] and the *Procedural and Conceptual Knowledge in Science* (PACKS) model [28]. *Understandings of Evidence* (UoE) [25], the insights a researcher uses to produce a cogent argument in support of a claim, are presented as the targeted learning goals.

A. Argumentation in inquiry

Argumentation is the process of reasoning systematically in support of an idea or theory, or “the uses of evidence to persuade an audience” [38]. As in science itself, it deserves a central and decisive role in science education [27,39–41], especially regarding scientific inquiry [38]. Even though the researcher may not yet know what peer criticism will receive, much thought and effort are invested in making the study’s claims as indisputable as possible and striving for optimal cogency of the argument in support of that claim. Convincing others of the validity of the claim includes describing the research procedures and methodological decisions as accurately and objectively as possible; justifying that the approach yields valid and reliable data; and demonstrating how these serve as evidence in support of the claim [42–44]. The researcher assesses alternative methods; analyses and interprets data; weighs evidence; considers various explanations for the observed phenomenon; and proactively defends the stated claims against potential criticism. All these actions are elements in the construction of a scientifically cogent argument [27,37,45]. Inquiry, from this perspective, can be interpreted as the construction of an optimally cogent argument that justifies the claim, i.e., the answer to the research question based on the data obtained [25,27].

B. The structure and content of a scientific argument

Gott and Duggan [27] adapted Toulmin’s model of argumentation [37] in which an argument consists of field-invariant as well as field-dependent elements, to the “field” of secondary science inquiry as shown in Fig. 1. The field-invariant elements include a *claim* based on *data* (facts or evidence) connected to each other through *warrants*: the reasoning defending the claim based on the data. Gott and Duggan elaborate this further in the context of school science: “the range of tried and tested methods of valid experimentation, as well as the substantive laws and principles, the complex network of interlocking theories that constitute the accepted body of scientific knowledge.” These warrants are further substantiated by *backings* which are considered, in the adapted model, the “detailed statements which underpin the data collection.” *Qualifiers* and *rebuttals* further strengthen the claim by setting limitations to its validity. As Toulmin points out, however: “If we ask about the validity, necessity, rigor or impossibility of arguments or conclusions, we must ask these questions within the limits of a given field...” [37] (p. 236).

In physics inquiry, the field-dependent content of an argument is provided by the PACKS model shown in

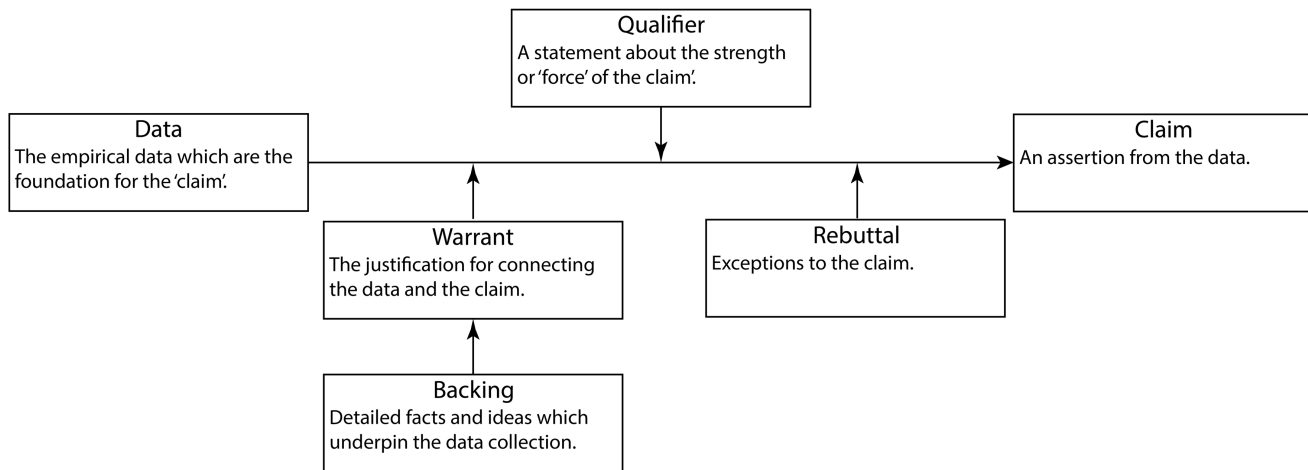


FIG. 1. Toulmin's argumentation model [37] adapted in Ref. [27] to secondary science inquiry.

Fig. 2. In this model, Millar *et al.* [28] link decisions made in various phases of an inquiry to four types of knowledge involving: (A) the (scientific) purpose of the inquiry; (B) the relevant content; (C) the required manipulative skills; and (D) the quality of scientific evidence. With regard to argumentation, e.g., PACKS knowledge type A influences students' interpretation of the task and thus influences the type of claim made by students [46]. While each of these knowledge types influences the decisions being made, knowledge type D is especially important in the construction of an argument in support of a claim [27,47–49].

Prominent elements in knowledge type D are the so-called *Concepts of Evidence* [46,49], which include concepts such as accuracy, range, and interval, which underpin the umbrella concepts of validity and reliability of data [48]. These Concepts of Evidence are the building blocks that enable us to construct, analyze, and judge a cogent account of the evidence [25,27,47]. In conceiving and conducting an inquiry, a researcher relies on insights in which these individual concepts acquire meaning through their relation to each other [25,50]. Pols *et al.* [25] explicated these insights as so-called *Understandings of Evidence* (UoE) of which the Concepts of Evidence are constitutive elements (see Table I for examples). These

UoE express not only what the quality criteria of scientific evidence are, and *how* they can be satisfied, but also *why* scientists adhere to them. For example, researchers are expected to repeat measurements, report means and spreads in the data, and if necessary apply a wide range of statistical techniques *because* it is understood that repeating a measurement naturally produces a range of values rather than a single one. In terms of Toulmin's model, the UoE provide field-dependent backings that contribute to the normative foundation on which the support for a claim is built. In an augmented Delphi study, Pols *et al.* validated a set of 19 UoE distributed over six inquiry phases as necessary and sufficient for evaluating evidence in physics inquiry at secondary school and first-year university level.

Pols *et al.* [25] constructed the "Assessment Rubric for Physics Inquiry" (ARPI) by specifying indicators for various attainment levels for all of these 19 insights on what allows to rate attainment of each UoE. They presented the UoE as the *learning goals in introductory activities directed at inquiry learning*. An appropriate selection of these UoE forms the set of learning goals for the teaching-learning sequence studied here. We explore an approach to developing selected UoE, and their contribution to students' regard for the quality of their inquiry and their ability to optimize it. The corresponding section of ARPI is used to

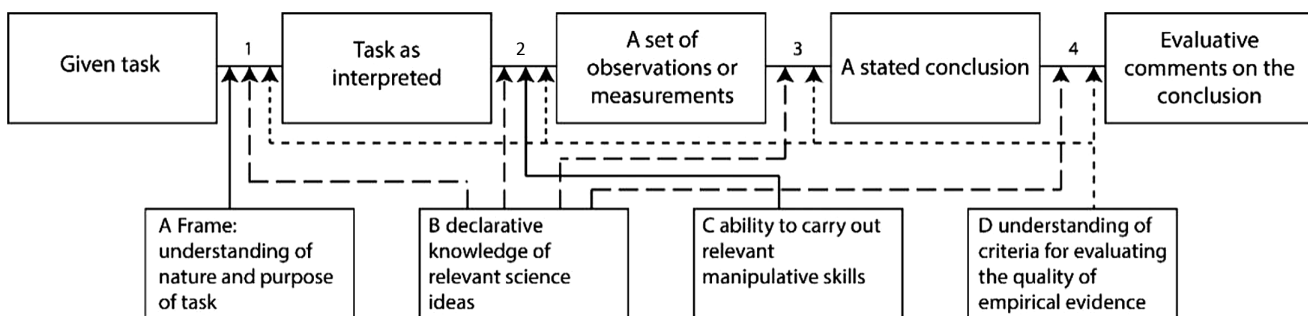


FIG. 2. The PACKS model [28] identifies various types of knowledge and their influence during inquiry tasks.

TABLE I. An overview of the UoE [25] that are selected as the learning goals for the teaching sequence with Concepts of Evidence [49] in bold.

Phase	UoE	The researcher understands that	This understanding is demonstrated by
2 Design	6	It is important to choose suitable instruments and procedures to get valid data with the required accuracy and precision .	Choosing and substantiating appropriate measuring instruments and procedures that provide the required reliability and accuracy of the dataset.
3 Method and Procedure	8	Measured values will show inherent variation and the reliability of data must be optimized, requiring repeated measurements .	Considering the number of repeated readings in terms of the required accuracy and/or available instruments and their sensitivity , adjusting the choice when needed and substantiating it.
	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.	Choosing and substantiating an appropriate and sensible measurement range and interval .
5 Conclusion and evaluation	14	A complete, clear, substantiated, and useful answer to the research question must be formulated.	Formulating a clear, substantiated, and unambiguous answer.
	16	The validity of conclusions does not go beyond the data available. Therefore, limitations to the validity of the claim should be expressed.	Specifying under what conditions the relationship /conclusion was established, discussing limitations.

monitor student progress in their ability to successfully engage in inquiry.

III. METHOD

A. Research design

Since we know of no previous empirical research into the integration of argumentation in inquiry, we had little to go on in terms of tested design principles or exemplary practices. We therefore elected to use educational design research [51,52] in this study. This research methodology is committed to simultaneously developing theoretical insights and practical solutions [53,54] through a combined study of the process of learning and the means that support it [51,55].

In the process of developing the teaching sequence, we derived learning goals from the list of UoE which we consider basic but essential for enabling students to undertake independent inquiry. These goals encompass fundamental insights into the underlying principles that guide actions we commonly anticipate young students to undertake, including practices like repeating measurements and using the broadest possible range for the independent variable. Subsequently, we curated a selection of activities from both existing literature [33–35] and our own instructional practices [26,36] that aligned with these identified goals. Informed by existing literature and our own expertise [15], we tentatively established design principles that in our view were necessary to integrate argumentation and inquiry, and sufficient to enhance learning from practical activities. Using these design principles, we then redesigned these

activities and tested them in the classroom. While each activity has been implemented and iterated individually, this test marked the initial iteration of a sequence involving all five activities as a cohesive whole.

As recommended [17,56–58], the feasibility of the design principles and the development of UoE in the teaching sequence were established in a small-scale, qualitative, in-depth, exploratory, single-classroom case study. It is directed at systematically exploring—through the qualitative instruments described in Sec. III E—and the realization of expected yields of applying the design principles derived in Sec. III C. This methodology is aimed at iteratively formulating and evaluating tentative design principles and concurrently constructing effective educational materials and activities on the basis of those principles, through closely monitoring, evaluating, and interpreting students’ responses and actions throughout the evolving design [51,54,59,60].

B. Participants and educational setting

The lessons were carried out by the first author at a regular Dutch school. It took place in his class of 23 students, aged 14–15 and in their final year of lower secondary education (grade 9), during their regular 50-minute physics lessons. With 10 years of teaching experience and a particular awareness of the challenges in teaching inquiry [61], he designed the sequence of activities.

Students’ work was graded, but merely handing in the work sufficed to obtain a passing grade (7 out of 10), in

order to reduce external pressure and emphasize the formative aspects of learning inquiry [62]. The students worked in largely fixed, self-selected teams of two students, or three to ensure no one worked alone. Due to illness, team G2 was canceled after activity 2 and does not appear in the data after that activity.

In Dutch lower secondary education, physics is mandatory and focuses on the development of scientific literacy and on preparing students for the optional science-based program in upper secondary education (Dutch: VWO) chosen by approximately 20% of the participating students. While there are national guidelines on the science content, there is no national exam at the end of lower secondary education [63,64]. Thus, teachers are to a large extent free to devise lessons and teach in the way they deem fit. For more details, see Ref. [15], where we argue that the sample is not exceptional, and findings are representative of many similar educational settings.

All names are fictitious, and all data were collected and treated in accordance with relevant ethical guidelines. All interventions, instruments, and collected data were in Dutch and have been translated by the authors. Practical aspects of the activities are described in more detail in Ref. [65]. Materials for all associated activities are open

source and available in English, Dutch, French, Spanish, and Basque [66].

C. Educational design

1. Educational aims

Table II summarizes for each of the five activities what the intended learning outcomes were in view of developing a deeper understanding of the scientific purpose of inquiry and PACKS type D (quality of evidence). Activities 1 and 2 focus on the development of the awareness that only the best possible answer suffices (UoE 14) and explore what “the best answer” precisely entails in science and a way to construct and defend such an answer. Activities 3 and 4 focus on the development of several common understandings used to gauge the quality of scientific evidence. We considered choosing suitable instruments and procedures (UoE 6); repeating measurements (UoE 8); and choosing a suitable range and interval for the independent variable (UoE 9) among the decisions to be undertaken by young students when engaging in basic inquiry. Understanding that the validity of conclusions does not go beyond the data available (UoE 16), in combination with the idea that the most informative conclusion should be drawn (UoE 14), is

TABLE II. An overview of the activities with the targeted UoE (Table I), number of teams participating, design principles (Sec. III C 2), and the data sources (Sec. III E).

Activity	Week	Data sources	Design principles	Targeted UoE	Main learning objective	Main activities
1. Pirate pendulum	1	i, iv, v	1–3	14	Developing the notion that in inquiry the best available answer is to be produced.	Students investigate the features of a pendulum in the context of a pirate movie stunt.
	2	ii, iii, iv	4–5			
2. Tricky tracks	3	ii, iii, iv	1–5		Distinguishing observation from interpretations and raising awareness of the need for argumentation in inquiry.	As in Ref. [35], students name what they observe in a given picture. Subsequently, they evaluate the reliability of evidence provided in a news article.
3. ISL	4	iii, iv	1–5	6, 9, 14, 16	Raising awareness of how the features of the dataset contribute to the quality of the data and the validity of the claims.	As in Ref. [36], students investigate the relation between body height and arm span in the context of a fair swimming competition.
4. Car crash barriers	5	iii, iv	1–5	8, 14	Developing the notion that variability in measurements is inevitable, finding an estimate of the true value, thus requires repeated measurements	As in Ref. [33], students investigate the relation between stopping distance and mass in the context of car safety.
5. NASA’s CRV	6	i, iv, v	1–3	Application of all of the above	Applying the acquired knowledge in an integrated way.	As in Ref. [67], students investigate the relation between a feature of a paper cone and its terminal velocity.
	7	i, iv, v				
	8	ii, iii	5			

TABLE III. The design principles and features derived from literature and the rationales, and expected returns.

No.	Label	Rationale	Expected returns
1	Guided inquiry	Offers a balance between autonomy and guidance	Students make their own methodological decisions, help is offered if requested
2	Reduction of knowledge demand	Ensures a focus on PACKS knowledge type D : the methodological decisions	Students know what to do and why (A), are not hindered by a lack of content knowledge (B), are able to work with the equipment (C)
3	A real life context	Shows the relevance of producing high quality answers	Students take answering their RQ seriously and mind the context in the discussions and in their answers
4	Productive failure	Offers time and opportunity for reflection, enables students to grapple with the ideas of evidence	The teacher presents ‘bad’ examples to initiate discussions on the quality of students’ decisions in their inquiry
5	Metacognitive tasks	Consolidates learning and strengthens the understandings of scientific criteria	Students formulate and apply personal but collectively agreed-upon ‘rules for doing proper investigations’ that are in line with the targeted UoE

essential in helping students understand how their methodological decisions influence the usefulness and trustworthiness of their conclusions and is therefore crucial in guiding students to make deliberate methodological decisions. Note that it would be naïve to think that each of these UoE is fully developed in a single activity. Although each activity focuses on the development of specific UoE, references to earlier targeted UoE are made throughout the lessons, thereby further developing and enhancing these UoE.

Based on these intended learning outcomes, suitable activities were chosen. Space limits us to elaborate on these here. For details on the activities see Ref. [65], the Supplemental Material [68] and the teacher manual [66].

2. Design principles

The activities were redesigned using a set of design principles that ought to enhance the integration of argumentation and inquiry and foster learning through practical activities. These design principles, described below, are linked to their intended student responses in Table III. In the Supplemental Material [68], more information on the activities and their links to the design principles can be found.

DPI guided inquiry. In *guided inquiry*, students are given a research question while they follow their own path to construct an answer [69]. In this way, for the population at hand, a balance between autonomy and guidance is expected to be provided. Autonomy is required to enable students to inquire into their problems in their own ways and learn from their successes and failures through reflection in sufficiently open activities [1,3,9,20,70–73]. Enough guidance, however, is required to overcome

problems that inexperienced students will otherwise experience as insuperable in inquiry [17,74].

This design principle is satisfied if students are observed to understand the research question or are able to formulate their own, and are able to devise a sensible way to answer it.

DP2 reduction of knowledge demand. Practical activities often present many (unnecessary) barriers that impede learning [22,74,75]. Summarized by Hodson [75], “[...] they have to understand the problem and the experimental procedure, assemble the relevant theoretical perspective, read, comprehend and follow the experimental directions, handle the apparatus, collect the data, recognize the difference between results obtained and results that should have been obtained, interpret the results, write an account of the experiment, and all the time ensure that they got along well with their partners.” Here, we evade the risk of cognitive overload by avoiding distracting details pertaining to PACKS type A–C [22,74,75]. We ensure that the activities are easily understood (A—scientific purpose), as simple as possible in terms of equipment (C—manipulative skills) and demand only previously acquired conceptual knowledge (B). If students still have questions related to these PACKS types, the teacher provides direct support and clarification so that the focus is on developing their understanding of the quality of evidence (D).

This design principle is satisfied if students are observed to have sufficient knowledge of pertinent theoretical concepts, measuring instruments, and methods to answer their research question. When it is satisfied, we expect students to focus on the quality of the answer and the evidence supporting it.

DP3 a real-life context. Practical activities rarely relate to students’ interests as these often lack a connection with the

“real world” [15]. For instance, it might be difficult for students to see how investigating the characteristics of a pendulum is relevant to them at all. However, framing the activity in real-world contexts [76–79] may make this connection. The real-world implementation of research findings can easily be understood to affect, for instance, people’s safety and thus make it easier to see why trustworthy data and conclusions are needed.

Understanding how the implementation of research findings is relevant in a given context is referred to as the *practicality of consequences*, number 87 in the list of Concepts of Evidence [80]. If truly adopted, the practicality of consequences provides students with the motivation to invest the required effort [15,26] and a need to extend their scientific knowledge [76]. However, it is known that students may see the context as window dressing or simply forget it [15,26,81]. An effort will have to be made for the context to be functional and taken seriously.

This principle is satisfied if students are observed to derive a motivation for obtaining convincing evidence and try to produce a useful answer to their research question in the given context.

DP4 productive failure. Unavoidably students make less-than-optimal decisions when given ownership and initiative in inquiry. To learn from these decisions, time and opportunity for feedback and reflection should be provided [82,83] where intervention by and negotiation with the teacher are essential [84]. As in other educational activities [85–87], we make productive use of decisions that students upon reflection regard to be suboptimal. We present these as “bad” examples that serve to address ideas pertaining to scientific evidence. Such discussions will “enable learners to grapple with the ideas of evidence affecting the quality of the work” [88]. When addressing these ideas systematically, students can become aware of the basis of decision making and apply their understanding to improve the quality of their data.

This principle is satisfied when students’ methodological decisions are observed to become the center of their attention in the activity, where students actively engage in becoming aware of their decisions, evaluating these, identifying their strong and weak aspects, and using these insights to direct their (future) decisions.

DP5 metacognitive tasks. While the activities are meant to be “open” in the sense that students devise their own procedures, they are “closed” in that all are meant to develop the same targeted UoE. Ultimately, students are enabled to use these UoE in forthcoming inquiries. Metacognition, often defined as “thinking about thinking” [89], is *what enables a student to apply a particular strategy in a similar but new context* [90]. Students’ metacognitive awareness of their understanding [91] is to be consolidated in metacognitive tasks that invite

them to reflect on, value, and organize the targeted knowledge [90,92]. Each activity ends with the metacognitive task to complete the following two sentences:

- 1) I learned in this activity that
- 2) I learned the following rules about doing inquiry: ...

Further metacognitive tasks are specified in the description of the activities given in the Supplemental Material [68].

This design principle is satisfied when students actively engage in reflecting on their newly obtained insights about establishing, critically evaluating, and defending evidence.

D. Aims and research questions

In Sec. III C 2, we specified design principles for practical activities that lower barriers to learning, enhance students’ critical evaluation of the quality of evidence, foster reflection on their own approach, and foster explicit understanding of scientific evidence. Investigating the activities’ outcomes guides an exploration of the role of argumentation in inquiry learning and of our theoretical understanding of that educational process. The corresponding first research question is

RQ1: Does our implementation of the selected design principles yield the expected returns in terms of students’ actions and attitudes?

If students do what we would like them to do, do they also learn what we intend them to learn? Or:

RQ2: To what extent do students attain the targeted UoE during the activities?

Finally, given what students learn in the activities, how does this affect their argumentation and their ability to plan and conduct a rigorous quantitative physics inquiry? Specifically:

RQ3: What progress is observed in students’ ability to engage independently in quantitative physics inquiry?

The combined answer to these questions provides the answer to the overarching research question:

What does integrating argumentation in teaching inquiry contribute to student understanding, critical attitude, and use of argumentation in doing quantitative physics inquiry?

E. Data sources

Among the main data sources are students’ written accounts of their work. Their work was stored in a portfolio, allowing them to evaluate what was done and learned. Their portfolios provided valuable information for this study as these contained:

- (i) Scientific graphic organizers—providing data for RQ1 and RQ3—are used in activities 1 and 5. A scientific graphic organizer is a prestructured lab journal where students report the essentials of an inquiry—research question, methods, instruments, essential theory, data represented in table and graph, and conclusions—without the necessity to write an extensive lab report [93,94]. Additional space for argumentation is provided.
- (ii) Student teams' written summary reports for activities 1 and 5 provide data that are triangulated with those of the scientific graphic organizer. No specific format was provided or required. Students were simply asked to report to the fictitious commissioners of their research and to detail whether they considered their findings to be reliable and why.
- (iii) Reflection forms pertaining to the metacognitive tasks in which students express their perceived learning gains which contribute to answering RQ2 about the students' developing UoE. Students' UoE are also elicited by asking them how they could have improved their earlier 'Pirates' inquiry in activity 1 as well as in their letter of advice to next year's students in activity 5.

These data were augmented in answering all research questions by:

- (iv) Audio recordings of all activities recorded using a microphone clipped to the teacher's shirt.
- (v) The teacher's field notes, providing a summary of each activity.

F. Data analysis

We specified the students' actions and attitudes (RQ1) that ought to be observed if the design principles were effectively implemented. We then compared these expectations with what actually happened in the five activities. Students' developing understanding of evidence (RQ2) was investigated by comparing students' statements expressed in metacognitive tasks involving the targeted UoE. Moreover, we established students' ability to operationalize the targeted knowledge. Using internal evaluation [60], students' enhanced ability to engage in independent quantitative physics inquiry was determined by comparing their initial and final understanding of scientific inquiry criteria and their attempts to adhere to these [25] (RQ3). Triangulation of the pre-post comparison with the data obtained during the various activities was used to identify particularly important aspects of the teaching sequence [60,95].

1. RQ1: Successful implementation of the design principles

Whether DP1-2 (guided inquiry and reduction of knowledge demand) were successfully implemented was established by verifying whether students produced the intended

output (data, graphs, and answer to the question). For DP3 (context), on the basis of audio recordings and their written answers to the research questions, we verified whether students made references to the context. We studied whether the implementation of DP1-3 elicited students' responses or instigated educational activities that can be expected to promote learning, e.g., a discourse in which a specific UoE is addressed. We analyzed whether addressing the weaknesses in students' approaches (DP4—productive failure) triggered discussions in which the issue at hand became the center of attention. For DP5 (metacognition), we established whether students produced solid answers—whether they indeed express, value, and organize their gained knowledge—in the reflective task and used these self-perceived insights to forward points of improvement pertaining to activity 1, the Pirates pendulum.

2. RQ2: Attainment of UoE

For each activity, we verified whether the students' perceived learning gains accorded with the intended learning outcomes and whether they applied the targeted knowledge. The overall development of UoE was determined by applying ARPI to students' work in activities 1 and 5, providing a broad, quantitative development pattern.

It is important to note that ARPI applies to PACKS knowledge type D (quality of evidence) and that the *minimum* level of attainment of a UoE is derived from the student's *actions* and *justifications* [25]. If a student makes a scientifically acceptable decision (e.g., repeats a measurement and reports the mean), the intermediate level (level 2, see Table IV) is ascribed as it may be the result of no more than rote learning. Level 4 is allocated only if a justification of this choice is provided as well. Since students tend to be brief in their explanations [96], we run the risk of underestimating a student's understanding. If a justification is lacking, this does not necessarily mean a student is unable to give it. While it has limitations, ARPI does provide a means to tentatively infer students' understanding from their actions.

3. RQ3: Students' ability to engage in quantitative physics inquiry

The development of students' ability to engage in inquiry independently was analyzed in terms of a selection of indicators specified for each activity. We compared activities 1 and 5 in terms of whether students spontaneously:

- (a) Construct the inquiry as an argument in support of a claim.
- (b) Take variability deliberately into account by repeating measurements, reporting means and spreads, addressing outliers.
- (c) Make deliberate choices in measuring instruments and procedures.
- (d) Make deliberate choices in data range and interval.

TABLE IV. The targeted UoE with five attainment levels. Indicators for three levels are provided. Intermediate levels are assigned when the lower level is outperformed but the higher level is not fully reached.

UoE	The researcher understands that	Attainment level		
		0	2	4
6	It is important to choose suitable instruments and procedures to get valid data with the required accuracy and precision.	Ignores options for selecting measuring instruments or procedures that would enhance data quality.	Considers options regarding instruments and procedures but fails to reach (independently) an optimal choice.	Makes an informed, substantiated, and acceptable choice between instruments and procedures so as to ensure optimally reliable and accurate data.
8	Measured values will show inherent variation and the reliability of data must be optimized, requiring repeated measurements.	Collects too few repeated measurements without substantiation or consideration of the quality of the dataset. Do not consider collecting further data at any stage.	Repeat 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.
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.	Measures inappropriate minimum, maximum, and/or in-between values.	Measured minimum, maximum, and/or interval are appropriate but lack substantiation.	Chooses and substantiates appropriate measured minimum, maximum, and interval.
14	A complete, clear, substantiated, and useful answer to the research question must be formulated.	Formulates an unclear and unsubstantiated answer which is insufficiently informative or insufficiently supported by the data.	Formulates a somewhat substantiated answer to the research question that is insufficiently informative or one where an explicit link between evidence and claim is missing.	Formulates a substantiated, optimally informative answer to the research question that is supported by the data available and presents the claim and evidence in a concise way.
16	The validity of conclusions does not go beyond the data available. Therefore, limitations to the validity of the claim should be expressed.	Does not discuss features and limitations that address the validity of the inquiry.	Discusses features and limitations to substantiate the validity of the inquiry and its outcomes, but inadequately or only partially.	Adequately substantiates limitations to the validity of the conclusion.

(e) Make their conclusions as informative and useful as possible by quantifying results and using data representations where appropriate.

(f) Apply a critical attitude toward their own approach and findings.

Criteria a, e, and f relate to understandings of the scientific purpose of inquiry that motivate finding the best available answer to the research question. Criteria b–d pertain to the basic choices that in every inquiry ought to be made but are rarely adequately made by students at this age. This set of criteria is therefore indispensable, yet tentative, in conducting an inquiry independently. We believe that it provides the foundational knowledge and skills necessary

for students to further develop their argumentation in inquiry. Students who spontaneously meet criteria a–f are applying the understandings and attitudes they are meant to develop.

The analysis is based on the students’ actions, decisions, and justifications reported in their scientific graphic organizer and written report. A qualitative comparison of activities 1 and 5 reveals salient patterns of development in students’ ability to engage in inquiry. Relating these qualitative findings to the quantitative data enables us to describe in more depth the relationship between this ability and students’ attainment of the targeted UoE. The analysis of statements and choices in students’ reports is limited in

scope. However, a further, in-depth qualitative analysis of classroom decision-making discussions among students and of consultations between students and teachers reveals some of the thinking behind the doing. It enables us to evaluate how the students' approach toward inquiry changed over time.

G. Reliability and validity

Studying one's own educational practice has the potential to bridge the research-practice gap [97], has high ecological validity [98], is accepted in both action research [99] and educational design research [51,53], and is advocated especially in the area of scientific inquiry [17,29,58]. Potential threats to data analysis bias [100,101] were minimized. The main data analysis (application of ARPI) and a significant part (30%) of secondary data analysis (student work) were carried out independently by the first author and a second teacher-researcher conversant with the teaching sequence. Rare cases of disagreement were discussed until a consensus was reached. Since only minor differences were found, the analysis is regarded to be sufficiently valid and reliable.

A main part of the data involved assigning attainment levels for students' UoE. However, students worked in small teams in all activities and eager and smart kids tend to take the lead in teamwork and whole-class activities. Strictly speaking, we therefore cannot regard these data as reflecting *individual* attainment levels. We consider this to be an acceptable tradeoff since individual work would have overly affected the authenticity of the lessons, where teamwork is common. To justify that the data do represent the whole class rather than the best performers, we include illustrative qualitative data pertaining to teams of varying assigned attainment levels.

IV. RESULTS

In presenting the data, we first explore whether each of the classroom activities has the outcomes that DP 1-5 are meant to effect (RQ1) using vignettes as illustrations (a detailed description of the implementation of the activities is provided in the Supplemental Material [68]). We then present students' reflections on their new insights (DP5),

describing and interpreting their words in terms of their alignment with selected UoE (RQ2). A comparison of data from the first and last activity establishes the overall, integrated progress in argumentation and its influence on the quality of students' inquiry (RQ3).

A. Design principles—realization and returns (RQ1)

Table V summarizes whether the use of the design principles rendered the intended yields in each activity. DP1, DP4, and DP5 were successfully implemented in all activities, while DP2 and DP3 were only partially successful in activity 5 and activities 3 and 4, respectively. These inferences are substantiated below. We provide quotes and vignettes that were selected for student responses and actions that are illustrative of what happened in class. When relevant to answering the research questions, we describe salient deviations from this general pattern. More qualitative details can be found in the Supplemental Material [66]. Without claiming objective proof for our interpretations of the data, we hope to provide them with credibility, comprehensibility, and traceability [102].

1. DP1 guided inquiry & DP2 reduction of knowledge demand

Throughout all activities, students understood what was required and prior knowledge sufficed to produce the intended products and outcomes. For instance, in activity 2, students understood the question "What do you observe?," interpreting it informally rather than scientifically. This, however, matched our intentions, no other prior knowledge was expected to be needed, and the informal interpretation sufficed. In activities 1, 3, 4, and 5 students' prior knowledge of physics content and measurement instruments sufficed to produce appropriate data with suitable procedures. In activity 5, assistance was asked and provided with regard to the use of accurate instruments, and adequate data were collected. At this point, only support in data analysis was not fully adequate as will be discussed further below. We conclude that DP 1 and 2 rendered the intended returns since students' actions were aimed at answering the intended research question which they understood and valued.

TABLE V. An overview of the design principles and the successful realization of these in each activity. Note that productive failure is not used in activity 5.

	DP 1: Guided inquiry	DP 2: Reduction of knowledge demand	DP 3: A real-life context	DP 4: Productive failure	DP 5: Metacognitive tasks
1. Pirate pendulum	+	+	+	+	+
2. Tricky tracks	+	+	+	+	+
3. ISL	+	+	±	+	+
4. Car crash barriers	+	+	±	+	+
5. NASA's CRV	+	±	+	×	+

2. DP3 real-life context

Inquiry contexts were found to be engaging and plausible to students and fostered a critical stance. For instance, when assessing the evidence for the existence of the Yeti in activity 2, they considered the quality of the evidence rather than (merely) preconceived opinion:

Teacher: Are you convinced that the Yeti exists?
 Julia (G11): No. I think it is a rather strange story. They found hairs, but it could also be of a wolf.

Throughout, the teacher was important in exploiting the context. For example in activity 4 (car crash barriers):

Teacher: Are you measuring correctly? There is some deviation in your measurements.
 Thim (G10): We measure correctly, but we will measure this one again, it probably is an exception.
 Teacher: But you measured it. Think about the car... If you just discard that measurement it might have severe consequences.
 Thim: Oh, yes, then it will end up in the canyon.

Thim decided to not discard the measurement but collect some more. Without prompting, another team challenged by the teacher decided to repeat *all* measurements. As it yielded the same spread in measurements, they complained:

Amy (G1): We measured again, but again our measurements deviate from each other.
 Teacher: What value will you report then?
 Amy: I would report this measurement as it is the only measurement showing up twice, and it is in the middle.

Amy's action—spontaneously taking responsibility for the lack of quality of the data by repeating all measurements—was exceptional but illustrative. (The inadequacy of her choice, of course, needs to be addressed [103,104] but is not our main concern here. The point is that she is *aware* of an issue that is relevant to her.) Even though the contexts were fictional and students sometimes needed help with interpreting their observations, they appeared motivated to obtain convincing evidence and prepared to be held accountable for the quality of their work.

3. DP4 productive failure

Activity 3 was performed fast, students collected body height and arm span in a mere 7 minutes, but often without considering proper procedures. They then did deem their methodological choices, such as not taking off shoes or not standing straight, inadequate but only in retrospect. Utilizing these unfortunate choices as “bad examples,”

the teacher highlighted the importance of choosing suitable procedures to get valid data. We show below how this contributed to students' attainment of UoE 6 (choice of suitable instruments and procedures).

The two vignettes from activity 4 presented above illustrate that the variability in students' measurements was at the center of their attention during the teacher-initiated talks. An important issue that emerged was students' expectation that a repeated measurement should render an identical result:

Eva (G5): We released the marble at the same spot, the cup was at the same position and the paper is fixed.
 Teacher: So you tried your utmost and still it does not yield the same result. Is that annoying?
 Masha (G5): Yes, you don't know what measurements you should use.
 Teacher: So, what would you do?
 Eva: Well, repeat it once again.

The large spread in measured values, a deliberate aspect of the set-up, seen as “bad” by the students, especially after the teacher's prompting, encouraged reflection on the measurement procedures, as our design intends.

In all activities' discussions, DP4 rendered intended returns in that exemplary decisions regarded as bad by the students themselves became the center of attention and instigated reflection on their approach and findings.

4. DP5 metacognitive tasks

After their final inquiry groups wrote a “letter of advice” for next year's students. Though ungraded, students' work showed metacognitive engagement and the realization of DP5:

G1: It is important to start with proper observations so that you already gain some knowledge about the experiment. Then it is important to measure as accurately as possible so that the results are reliable.
 G10: Think thoroughly before you start taking measurements, consider what you are going to do, and make sure that you understand the research. This will result in a proper research that has actual value for you as well. [...] It is not about finishing as quickly as possible, it is about whether you devise a proper research.

DP5 was implemented successfully: data show that students engaged thoroughly in reflecting on what they had learned. In each activity perceived, learning gains aligned well with the intended outcomes. The students, however, remained presently often unable to specify their

TABLE VI. Students’ attainment levels for activities 1 (Pre) and 5 (Post). Shown is the number of teams per competence level for each UoE [25] on a five-point scale from lowest (0) to highest (4), on average in scientific graphic organizer and letter. Class average level in gray, and deviations larger than 0.5 in the mean score are darker gray. Number of teams whose UoE could not be determined in “no score” column.

Phase	UoE	The researcher understands that:	Activity	No score	0	1	2	3	4
Design	6	It is important to choose suitable instruments and procedures to get valid data with the required accuracy and precision.	1	0	10	1	0	0	0
			5	0	2	3	1	1	3
Methods & procedures	8	Measured values will show inherent variation and the reliability of data must be optimized, requiring repeated measurements.	1	0	2	1	8	0	0
			5	0	0	1	5	2	2
	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.	1	2	3	0	3	2	1
			5	0	0	0	9	1	0
Conclusion & evaluation	14	A complete, clear, substantiated and useful answer to the research question must be formulated.	1	1	3	3	0	4	0
			5	0	3	2	4	1	0
	16	The validity of conclusions does not go beyond the data available. Therefore limitations to the validity of the claim should be expressed.	1	0	2	4	4	1	0
			5	1	3	2	3	0	1

insights at a level of abstraction that would be required to transfer these insights to future inquiries.

Development of UoE (RQ2). Table VI shows the result of applying ARPI to the scientific graphic organizer and reports in activities 1 and 5. Though only semiquantitative, it provides a global view of development (RQ2), further substantiated in the next section. Average attainment improved in UoE 6 and 8 suggesting potential progress in students’ understanding. Low scores here were mostly caused by the absence or brevity of explanations and justifications. While in UoE 9 the average level did not change, all students spontaneously used the appropriate range and interval in activity 5—but failed to fully explain why they did and what they did. There is no observable change in the average attainment of UoE 14 and 16. While the qualitative data do suggest that attainment in UoE 14 was present, students’ difficulty with numerically analyzing the data caused them to draw partly unsubstantiated conclusions reducing their attainment level.

B. Students’ ability to engage in inquiry independently (RQ3)

We compare activities 1 and 5 in terms of the six criteria a–f specified in Sec. III “data analysis to describe students” progress in engaging in independent inquiry (RQ3). Salient data from other activities augment the

findings. In this section as before, illustrative vignettes and quotes are used to provide our interpretations with credibility, comprehensibility, and traceability, with more details in the Supplemental Material [68].

1. Criterion a. Construct the inquiry as an argument in support of a claim

Activity 1: All teams reported in their letter to the stunt coordinator what was done. For instance, team G6 reported that it had investigated the influence of the type of rope on the swing time (Fig. 3). Only 4 teams reported how they investigated, while 9 out of the 11 letters omitted information needed to verify their claim. As illustrated in Fig. 3 data and claims were seemingly presented as uncontested and producing a convincing argument as therefore unnecessary.

Activity 3: At this stage, a majority (8 out of 10 teams) of students’ reflections referred to concerns about the reliability of conclusions, for example:

G8: In this activity, we learned that in constructing a conclusion you have to ensure that your information is reliable. The conclusion should answer the research question. That the thing you investigate is adequately tested. You have to explain how you arrived at your answer. You have to ensure that the conclusion is useful to others.

Dear stunt coordinator,

(What) For the stunt, we investigated different types of rope. **(Claim)** We reached the conclusion that the difference is not big, but even this small difference can be very important in the timing of the jump. The thinner and lighter the rope is, the longer the [swing] takes. **(Recommendation)** It is therefore necessary to carefully consider which rope type best suits the jump. We hope to have helped you with this and for further details please see the research sheet [scientific graphic organizer].

FIG. 3. Team G6's letter to the stunt coordinator, in bold our analysis pertaining to the elements of the letter.

Activity 4: Students began to incorporate their conclusions and also attempt to quantify, justify, and explain these.

- G3: The stronger a crash barrier, the shorter the stopping distance: If the crash barrier is four [cups] stronger, the stopping distance is roughly halved. You have to do additional measurements [on] heavier cars to find the precise relation. The conclusion should not be based on the average value, but the maximum distance.
- G10: We have conducted an experiment with cups and marbles. We have measured 4 times, first with a single cup, repeated it with an additional cup stacked, and so on up to 4 stacked cups. Each measurement is repeated 5 times for the most reliable result. We did this for a specific marble and cup. In reality, a car might be heavier when filled with luggage. One has to pay attention to that as well. Each time another cup was added, the stopping distance halved. So, the stronger the crash barrier, the shorter the stopping distance. [...] Our advice is to repeat the test with real cars for a more reliable result, the results might be different as we just used marbles.

Since students were not prompted to include these arguments, we infer that they had come to understand that the answer to the research question in inquiry requires a supporting argument based on the data. Without that understanding, it is highly unlikely that they would spontaneously have tried to provide these arguments.

Activity 5: The letter of G6 in Fig. 4 illustrates how all teams provided information about what was investigated and how at the end of the teaching-learning sequence. Nine of the ten teams provided details, including warrants

and backings that allow for an external assessment of the inquiry's quality. These students presented their measurements as an explicit—albeit limited—substantiation of their answers to their research questions. They thus constructed their inquiry as an argument in support of a claim.

2. Criterion b. Take variability into account by repeating measurements, reporting means and spreads, addressing outliers

Activity 1: Most teams repeated measurements routinely 3 times here. As this is the standard but unsubstantiated practice in secondary school physics, we allocated the intermediate attainment level (see Table VI). Three teams (G7, 10, and 11) reported repeating measurements, and no team considered data variability or outliers.

Activity 4: The above data on the returns of DP3 (context) and DP4 (productive failure) illustrate students' struggle in coming to terms with the inevitability of variability in measurements and the need for repeated measurements. Their learning is reflected in the rules they formulated eventually:

- G3: You have to take the purpose of the activity into account when considering whether you use only the average or all measurements, even those that deviate. By repeating measurements, the result becomes more precise.
- G4: Take as many measurements as possible to obtain a better idea of the outcomes of the experiment.

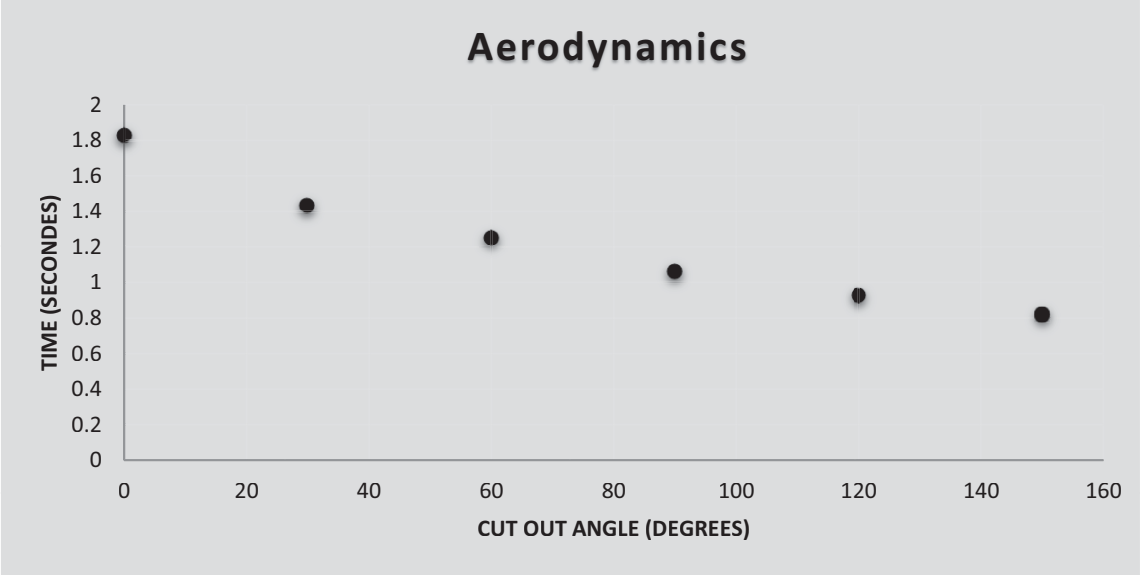
Activity 5: Now, all teams chose to repeat each measurement, usually 5 times. Two teams (G1 and G8) reported taking outliers into account by taking more measurements when they saw a clear deviation. Seven teams provided all repeated measurements in their letter to NASA, three used a

Dear NASA,

(What) We have investigated the influence of aerodynamics on the fall speed of the CRV.
(How) We have made a miniaturization of the CRV by making cones from circles with a diameter of 15.7 cm. To get an accurate measurement, we made 6 cones, each with a different cut-out angle (0°, 30°, 60°, 90°, 120°, 150°). The greater the angle cut, the more streamlined the cone is. We dropped the cones 5 times from the same height so that the measurement is as precise as possible (**substantiation of choice**). We recorded the fall time with a stopwatch. We took an average of the 5 measurements and made a graph. **(Claim)** From this you can conclude that the more streamlined the cone (CRV) is, the faster it will fall down. If you take 120° more from the circle, the cone falls twice as fast.

(Backing) Our results are very reliable because we made many different cones to get a clear relation. Also, we have dropped the cone several times to increase reliability of the measurements. There was another team investigating the influence of aerodynamics on the fall speed of the CRV. We compared our results with the results of the other group and found that the results match. This makes the measurements even more reliable.

(Limitation) We recommend to further investigate the factors that influence the fall speed. These factors all together ultimately determine the fall speed.



Cut Out Angle (Degrees)	Time (Seconds)
0	1.8
30	1.45
60	1.25
90	1.05
120	0.95
150	0.85

FIG. 4. Team G6’s letter to NASA, in bold our analysis pertaining to the elements of the letter.

graph to display only the average value. Five teams linked the number of repeated measurements to the reliability of results, but often obscurely:

- G6: We have dropped the cone several times (5×) to increase the reliability of the measurements.
- G8: At some heights more than 5 repeated readings were taken due to aberrant readings ...the results are still not 100% correct, because we manually used a stopwatch.

In peer feedback teams that repeated 3 times were advised to take more readings next time:

- G10: Three is the bare minimum, you should take at least five repeated readings as you also have reaction time.

The upward trend in ARPI score on UoE 8 (variation in measurements) seen in Table VI between activities 1 and 5 suggests an enhanced awareness of the importance of repeated measurements and how to report them.

These qualitative data not only confirm that trend but also show that students are not yet equipped to choose and justify the number of repeated measurements adequately.

3. Criterion c. Make deliberate choices in measuring instruments and procedure

Activity 1: All students used the readily available instruments: a ruler and stopwatch. Most teams timed half a period and made no attempt to optimize this procedure. They did not search for (more) suitable instruments or procedures nor consulted the teacher about this.

Activity 5: Most teams now showed awareness of the need to produce accurate and precise data by optimizing the choice of instruments and procedures. Most actively ensured control and measurement of the falling distance. Half the teams actively optimized time measurements by, e.g., filming the falling cone together with a stopwatch, and analyzing the images in slow motion. Two teams provided a reasoned substantiation, G1 most elaborately:

G1: When you measure with a stopwatch, you have to deal with your own reaction time, so your measurements will always deviate a little from the truth. Because the light gates accurately measure the fall time of the cone in milliseconds, you can be sure that the measurements are reliable.

Still, five teams did not change their approach to measuring the time between activities 1 and 5. Four of these did provide recommendations to improve this but only in hindsight, so ARPI score 1 was assigned. We infer an enhanced awareness of the necessity to choose appropriate instruments and devise proper procedures. However, it is not always done in a timely manner and at the appropriate time (e.g., during the research design stage).

4. Criterion d. Deliberate choices in data range and interval

Activity 1: About half the teams displayed data range and interval issues (e.g., choosing too few independent values or values from an insufficient range for a relationship to emerge). Only team G11 specified its choices:

G11: We used a rope length of 50, 100, and 150 cm to verify whether the swing time doubled when the length of the rope doubles. Unfortunately, this is not the case.

Activity 5: Now all teams chose a range and interval that allowed a pattern to be revealed. Most frequently the range included six different values. Since a justification of the choice was absent, however, nearly all teams were assigned intermediate level. While the changes in their

unprompted choices suggest that these are made more deliberately, that deliberation is not translated into an explicit justification.

5. Criterion e. Make their conclusions as informative and useful as possible by, where possible, quantification of results and using data representations such as tables and graphs

Activity 1: Only team G11 attempted to establish a quantitative relation. However, this was mostly acceptable since several correctly concluded that the relation they investigated did not exist. Only two teams warranted their conclusion by providing a summary of the data:

G1: We did not find larger differences in swing time as with an angle of 10 degrees the average swing time was 0.60 s and at an angle of 40 degrees 0.58 s.

Team G3 provided their measurements in a table; no other group used a table or graph.

Activity 5: Seven teams provided a quantitative statement as their conclusion (though not always in accord with theory). All teams presented their data using a graph or table, and half the groups using both a graph and a table. Their attempts at quantification are generally not very successful from a scientific perspective, as they lack data analysis skills. The low ARPI scores in Table VI therefore suggest little progress during the teaching-learning sequence. The information on progress, however, is less about what they accomplish and more about their effort:

G4: If the height becomes twice as high, the fall time will be about 1.8 [times] as long. This only applies from 80 cm.

G6: We conclude that the more streamlined the cone (CRV) is, the faster it will fall. If you cut out a section of 120 degrees from the circle, the cone falls twice as fast.

These examples illustrate their attempts to quantify and do show an increased awareness that the inquiry ideally ought to result in a conclusion that expresses a quantitative relationship between the investigated variables.

6. Criterion f. Critical attitude towards own approach and findings

Activity 1: By their own reflective admission the students showed no critical stance toward their approach during data collection. They were more self-critical in hindsight, when they were asked to consider being the stuntperson, facing the consequences of the decisions taken in making the stunt which are based on their own data. They deflected responsibility in statements such as “we were not given appropriate equipment.”

Activity 5: In the last activity, the students showed an enhanced critical stance toward their own approach by, e.g., increasing the number of repeated measurements (all except G9 and 11), and deliberately choosing more accurate instruments (G1, G5, G9, and G10). Rather than deflecting responsibility, the various teams presented shortcomings of their approach in their recommendations. Their reflection on the quality of their approach aligned with a scientific perspective.

V. DISCUSSION

Below, we analyze to what extent integration of argumentation in inquiry resulted in the development of selected UoE and how that contributed to students' ability to engage in inquiry.

A. Evaluation of the implementation of design principles (RQ1)

DP1 (guided inquiry) was meant to balance guidance for students to maintain progress in their inquiries with autonomy to use and evaluate their own ideas. In all activities, students progressed smoothly and yet used their own approaches. They explored, guided by the teacher, the quality of their answers and justifications. We conclude that DP1 was implemented successfully.

DP2 (reduction of knowledge demand) aimed to eliminate distracting theoretical and procedural issues in order to provide enough time and energy to think and talk in class about how to obtain the best possible answer. Activities 1–4 were clearly simple enough for the students content-wise. This allowed them to develop nontrivial concepts of evidence such as fair test, variability, repeatability, and outliers. They did experience problems in interpreting the data in activity 5, but we attributed this to their own increased demand for the quality of scientific evidence rather than ignorance. We see this cognitive need for PACKS type B (content knowledge) and C (manipulative skills) therefore as a further success of the activities. DP2 was successfully implemented.

DP3 (context) aimed to motivate students to invest enough time and effort in their inquiry through consideration of the *practicality of consequences* that would result from actually applying their research findings. The teacher's references to the contexts helped students to attach meaning to the concepts and understandings of evidence. While students attached relevance to the fictitious contexts and took them seriously, the contexts did not cause them to spontaneously strive for optimal scientific quality. However, even the briefest reminder of the "Pirates" activity made students reconsider the adequacy of their approach later on. While only partially realizing intended yields, DP3 contributed in important ways to the integration of argumentation in inquiry.

DP4 (productive failure) may easily be misunderstood. It is *not* about telling students what they did wrong, but about utilizing what *they* regard as a mistake, by having them reflect on it, so as to promote learning. Note, for example, that in activity 4, the students' feeling that they ought to produce data with less variation was in fact not a scientific failure at all and used to discuss the concepts of evidence of natural variability in measurements. As our data show, DP4 was successfully exploited in activities 1–4.

DP5 (metacognition) is indispensable in any teaching-learning activity that integrates argumentation. Arguments come into play only if claims are questioned, which requires reflection, consideration of past statements and actions, contemplating their implications, and considering alternatives. The crux of the matter is not the use of these activities *per se*, but designing them in such a way that students (and teachers) see their relevance and experience their value. We believe that in all activities the exchanges among students and between students and the teacher demonstrate that this was the case. Students' ability to specify rules for doing proper research and to use these in recommending improvements to earlier inquiries show that these activities satisfied design intentions.

This discussion on the implementation of DP1-5 consecutively, as if the contribution of each can be isolated from the others, is a simplification meant to highlight specific attributes of the teaching-learning sequence. In the classroom, the design principles actually interact and their combined implications are experienced.

B. Evaluation of the development of targeted UoE (RQ2)

Where students chose the first method at hand to measure time and distance without consulting the teacher in the first activity, half of the teams consulted the teacher in devising a reliable method in the last. Their spontaneous request for help we believe indicates a readiness for the development of knowledge of type B (content) and C (manipulative skills) and is based on attaining understanding *UoE 6* (choice of instruments and procedures).

UoE 8 (variation in measurements) highlights the importance of not relying on a single reading or blindly following a previously prescribed "rule." Students initially repeated measurements without reason, as seen in activity 1 [26]. In activities 1 and 4, they were found to hold the naïve belief, often reported in the literature, that repeated measurements should yield identical results [103,105–107]. However, the metacognitive reflections in activity 5 show that all developed a deeper understanding of *UoE 8* and more deliberately chose a larger but fixed number of repeated measurements. They did not produce explicit reasons to substantiate the sufficiency of this number nor did they relate it to the variation in the measurements. Indeed, developing more than an intuitive understanding of "enough" repeated readings at this age is challenging [104],

as it requires the ability to quantify variation and calculate and interpret measurement uncertainty [107].

Students initially often obtained results of little value because they experimented only with small variations in the independent variable, i.e., the pendulum length in activity 1. Indicative of *UoE 9* students' choices of range and interval improved as all chose a range and interval that revealed a pattern at the end. However, a justification for the choice remained absent.

ARPI scores in activity 1 were relatively high. We attribute this to students drawing correct conclusions about the nonrelationships they established. They were correct but simple, expressed in statements like "the results did not differ much." Our data show that, over the course of the sequence of activities, students' growing awareness that a conclusion must be as informative as possible (*UoE14*). *UoE 14* provides them with reasons for investing the time and energy needed to design and conduct a rigorous inquiry. Drawing conclusions that were quantitative in nature, however, remained difficult due to a lack of knowledge in data analysis.

UoE 16 (validity of conclusions) explains why a maximum range of the independent variable and size of the sample should be chosen and clarifies the relevance of specifying the conditions under which the results have been obtained and that explicating them contributes to the credibility of the study and its findings. It problematizes extrapolation and interpolation. The data show students' increasing awareness of providing specific information on data collection. They did not improve in specifying the limitations of their inquiry, and thus there is room for further growth in this area.

C. Evaluation of progress in students' ability to engage in inquiry (RQ3)

Activity 5 data indicate that students attempted to perform and report their inquiry with the purpose of producing a scientifically convincing argument in support of a claim (criterion *a*). Though limited in extent and quality, students included *warrants and backings* in their letters while none were present in their work in activity 1.

Students' enhanced *UoE* ensured that during the planning of the inquiry they recognized the methodological choices they had to make. The data show that they considered more deliberately what scientifically acceptable decisions are (criteria b–d). Even students who did not make scientifically more desirable decisions during the design did specify such improvements in retrospect.

Students showed an increased awareness of what is expected in drawing and specifying scientific conclusions in inquiry. The majority tried to make their conclusion informative by quantifying their results and substantiating it by presenting the data (criterion *e*).

Students' initial approach can be described as "controlled chaos," in which they seemingly unthinkingly

gathered data to quickly "get the job done." Many teams progressed toward a more "systematic" approach in which they considered different methods and procedures using their acquired understanding. Without being instructed to do so, they started to use "rules" and procedures based on what they themselves formulated during the teaching-learning sequence to obtain reliable and valid data. While limited and largely remaining implicit, these findings indicate an increased critical attitude and students' consideration of the question "what is the best next step in the inquiry within the existing constraints?" (criterion *f*).

Our criteria for the use of argumentation in inquiry have been met to the extent that students started taking responsibility for the quality of their investigation and applied their acquired understanding, but not to the extent that they justified each decision. The step from fostering students' searching for justifiable actions to enabling them to actually scientifically justify these actions is one of the many next steps in teaching inquiry. We have merely described a potential first step in enticing students to make their ideas about constructing evidence explicit to others in the ways accepted in science.

VI. CONCLUSIONS AND FUTURE RESEARCH

We investigated how focusing on argumentation contributes to enabling young (grade 9), inexperienced students to engage in independent scientific inquiry. The design principles used in the sequence of activities enabled students to engage in inquiry purposefully and use argumentation to strive for the best possible answer to their research question. They began to consider several core characteristics of scientific evidence and aimed to construct scientifically convincing arguments—guided by the *UoE* addressed in the teaching-learning sequence—without being instructed to do so in the final inquiry. Although operationalizing the *UoE* in an integrated way remained difficult and their research did not improve much if judged by the traditional technical standards, students showed increased awareness of research quality. They developed a better understanding of *why* they are expected to try and meet these standards and showed a readiness for learning *how* to do so.

We previously established a set of learning goals (the *UoE*) for integrating argumentation with inquiry [25] and created ARPI for assessing the attainment of these. Here we extended and further clarified the framework for integrating argumentation in inquiry by linking the *UoE* and the *Procedural and Conceptual Knowledge in Science* model to Toulmin's argumentation model. The *UoE* are identified as central among the field-dependent elements that determine the cogency of a scientific claim. We have now developed and tested design principles, viable and feasible in this setting, for the first steps students take in attaining these learning goals. Engaging in argumentation is promising to contribute in important ways to students' ability to

engage in inquiry independently. We intend to further develop theatrical implications and evaluate their practical implications and value.

In that respect, testing the teaching-learning sequence in other settings is in order. The influence of and support for the teacher require further research as teachers are often ill-equipped to teach scientific inquiry [4,21,108,109] and integrating argumentation implies further demands. While collaboration between school subjects in developing inquiry is indispensable, more research is needed also on effective knowledge transfer [110–112].

We believe that our findings also have implications for recent efforts that have been made in teaching measurement uncertainty and gauging students' understanding thereof [104,113,114]. In those studies, the emphasis seems to be on students' ability to make sense of data and apply different (mathematical) concepts. The findings in this study seem to indicate that these concepts are better learned when students have a better understanding of the scientific purpose of experimental physics. Concepts regarding measurement uncertainty can then be used by students to quantify to what extent *the best answer* to the research

question has been produced, and how good that answer actually is. Measurement uncertainty then becomes a tool in their extensive argumentation toolbox to guide their inquiry.

In authentic, more complex inquiries than those of the teaching-learning sequence, the PACKS knowledge types inevitably interfere while multiple UoE are to be applied simultaneously. Further research is needed to effectively address the challenge of developing other UoE—in what order, at what level, and to what extent—and the integration of PACKS knowledge types. We think integrating argumentation in inquiry is long overdue and a promising avenue for research into learning to do inquiry.

ACKNOWLEDGMENTS

This work is part of a research program for teachers financed by the Netherlands Organisation for Scientific Research (NWO) [Grant No. 023.003.004]. The materials have been translated into various languages [66] with the help of the Erasmus⁺ program of the European Union [Grant No. 2018–1-FR01-KA201-048202].

-
- [1] D. Hodson, Learning science, learning about science, doing science: Different goals demand different learning methods, *Int. J. Sci. Educ.* **36**, 2534 (2014).
 - [2] J. Kozminski *et al.*, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum*, edited by A. A. o. P. Teachers (2014), p. 29.
 - [3] A. Hofstein and P. M. Kind, Learning in and from science laboratories, in *Second International Handbook of Science Education*, edited by B. Fraser, K. Tobin, and C. J. McRobbie (Springer, Dordrecht, The Netherlands, 2012), pp. 189–207.
 - [4] V. N. Lunetta, A. Hofstein, and M. P. Clough, Learning and teaching in the school science laboratory: An analysis of research, theory, and practice, in *Handbook of Research on Science Education*, edited by N. Lederman and S. K. Abell (Lawrence Erlbaum Associates, Hillsdale, NJ, 2007), pp. 393–441.
 - [5] A. Hofstein and V. N. Lunetta, The laboratory in science education: Foundations for the twenty-first century, *Sci. Educ.* **88**, 28 (2004).
 - [6] R. Millar and J. Osborne, *Beyond 2000: Science Education for the Future* (King's College London, School of Education, London, 1998), <https://www.nuffieldfoundation.org/wp-content/uploads/2015/11/Beyond-2000.pdf>.
 - [7] Next Generation Science Standards, Next generation science standards: For states, by states. Appendix D: All standards, all students: Making the Next Generation Science Standards accessible to all students (2013).
 - [8] I. Abrahams and M. J. Reiss, Practical work: Its effectiveness in primary and secondary schools in England, *J. Res. Sci. Teach.* **49**, 1035 (2012).
 - [9] N. G. Holmes and C. Wieman, Introductory physics labs: We can do better, *Phys. Today* **71**, No. 1, 38 (2018).
 - [10] R. Millar, J. F. Le Maréchal, and A. Tiberghien, Mapping the domain: Varieties of practical work, in *Practical Work in Science Education—Recent Research Studies*, edited by J. Leach and A. Paulsen (Roskilde University Press/Kluwer, Roskilde/Dordrecht, The Netherlands, 1999), pp. 33–59.
 - [11] R. Cross, Rolling and sliding down an inclined plane, *Phys. Teach.* **61**, 568 (2023).
 - [12] F. Pols, The sound of music: Determining Young's modulus using a guitar string, *Phys. Educ.* **56**, 035027 (2021).
 - [13] A. Gkourmpis, Discovering the laws of gases with a manometer based on Arduino, *Phys. Teach.* **61**, 500 (2023).
 - [14] F. Boczianowski and B. Priemer, The spinning toilet brush—a classroom experiment on the mechanical equivalent of Joule's heat, *Phys. Educ.* **58**, 065012 (2023).
 - [15] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. de Vries, What do they know? Investigating students' ability to analyse experimental data in secondary physics education, *Int. J. Sci. Educ.* **43**, 274 (2021).

- [16] Z. Kanari and R. Millar, Reasoning from data: How students collect and interpret data in science investigations, *J. Res. Sci. Teach.* **41**, 748 (2004).
- [17] D. Hodson, A critical look at practical work in school science, *Sch. Sci. Rev.* **70**, 33 (1990).
- [18] R. Tasker and P. Freyberg, Facing the mismatches in the classroom, *Learning in Science: The Implications of Children's Science* (Eric, Portsmouth, 1985), pp. 66–80.
- [19] F. Lubben and R. Millar, Children's ideas about the reliability of experimental data, *Int. J. Sci. Educ.* **18**, 955 (1996).
- [20] N. G. Holmes and C. Wieman, Examining and contrasting the cognitive activities engaged in undergraduate research experiences and lab courses, *Phys. Rev. Phys. Educ. Res.* **12**, 020103 (2016).
- [21] I. Abrahams and R. Millar, Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science, *Int. J. Sci. Educ.* **30**, 1945 (2008).
- [22] E. van den Berg, The PCK of laboratory teaching: Turning manipulation of equipment into manipulation of ideas, *Sci. Educ.* **4**, 74 (2013).
- [23] C. F. J. Pols, A physics lab course in times of COVID-19, *Electron. J. Res. Sci. Math. Educ.* **24**, 172 (2020).
- [24] J. Osborne, Teaching scientific practices: Meeting the challenge of change, *J. Sci. Teach. Educ.* **25**, 177 (2014).
- [25] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. de Vries, Defining and assessing understandings of evidence with assessment rubric for physics inquiry: Towards integration of argumentation and inquiry, *Phys. Rev. Phys. Educ. Res.* **18**, 010111 (2022).
- [26] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. de Vries, "Would you dare to jump?" Fostering a scientific approach to research in secondary physics education, *Int. J. Sci. Educ.* **44**, 1481 (2022).
- [27] R. Gott and S. Duggan, A framework for practical work in science and scientific literacy through argumentation, *Res. Sci. Technol. Educ.* **25**, 271 (2007).
- [28] R. Millar *et al.*, Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance, *Res. Pap. Educ.* **9**, 207 (1994).
- [29] A. Hofstein, The role of laboratory in science teaching, and learning, in *Science Education*, edited by K. S. Taber and B. Akpan (Springer, Dordrecht, 2017), pp. 357–368.
- [30] R. Roberts and P. Johnson, Understanding the quality of data: A concept map for 'the thinking behind the doing' in scientific practice, *Curric. J.* **26**, 345 (2015).
- [31] R. Driver, P. Newton, and J. Osborne, Establishing the norms of scientific argumentation in classrooms, *Sci. Educ.* **84**, 287 (2000).
- [32] R. Watson, J. R. Swain, and C. McRobbie, Students' discussions in practical scientific inquiries, *Int. J. Sci. Educ.* **26**, 25 (2004).
- [33] S. Farmer, Real graphs from real data: Experiencing the concepts of measurement and uncertainty, *Sch. Sci. Rev.* **346**, 81 (2012).
- [34] A. Mooldijk and E. Savelsbergh, An example of the integration of modeling into the curriculum: A falling cone, in *Proceedings of the GIREP: Physics Teacher Education Beyond* (GIREP, Barcelona, 2000), pp. 625–628.
- [35] N. G. Lederman and F. Abd-El-Khalick, Avoiding de-natured science: Activities that promote understandings of the nature of science, in *The Nature of Science in Science Education* (Springer, New York, 1998), pp. 83–126.
- [36] C. F. J. Pols, The Vitruvian man: An introduction to measurement and data analysis, *Phys. Teach.* (to be published).
- [37] S. E. Toulmin, *The Uses of Argument* (Cambridge University Press, Cambridge, England, 2003).
- [38] G. J. Kelly, Discourse practices in science learning, and teaching, in *Handbook of Research on Science Education*, edited by N. Lederman and S. K. Abell (2014), pp. 321–336.
- [39] P. Newton, R. Driver, and J. Osborne, The place of argumentation in the pedagogy of school science, *Int. J. Sci. Educ.* **21**, 553 (1999).
- [40] S. Erduran and M. P. Jiménez-Aleixandre, *Argumentation in Science Education. Perspectives from Classroom-Based Research* (Springer, Dordrecht, 2008).
- [41] J. Osborne, The 21st century challenge for science education: Assessing scientific reasoning, *Think. Skills Creat.* **10**, 265 (2013).
- [42] N. Oreskes, The scientific consensus on climate change: How do we know we're not wrong? in *Climate Modelling* (Springer, New York, 2018), pp. 31–64.
- [43] A. F. Chalmers, *What is This Thing Called Science?* (Hackett Publishing, Indianapolis, IN, 2013).
- [44] American Psychological Association, *Publication Manual* (American Psychological Association Washington, DC, 1983).
- [45] S. Woolgar and B. Latour, *Laboratory Life: The Construction of Scientific Facts* (Princeton University Press, Princeton, NJ, 1986).
- [46] R. Millar, Student's understanding of the procedures of scientific enquiry, in *Connecting Research in Physics Education with Teacher Education*, edited by A. Tiberghien, E. L. Jossem, and J. Barojas (International Commission on Physics Education, 1997), pp. 65–70.
- [47] R. Gott and R. Roberts, *Concepts of Evidence and their Role in Open-Ended Practical Investigations and Scientific Literacy; Background to Published Papers* (The School of Education, Durham University, UK, 2008).
- [48] R. Gott *et al.*, Research into Understanding Scientific Evidence 2003–2018 [cited 2019]; available from <http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm>.
- [49] R. Gott and S. Duggan, Practical work: Its role in the understanding of evidence in science, *Int. J. Sci. Educ.* **18**, 791 (1996).
- [50] R. White and R. Gunstone, *Probing Understanding* (Routledge, London, 1992).
- [51] J. Van den Akker *et al.*, *Educational Design Research* (Routledge, Abingdon, Oxon, 2006).
- [52] P. Bell, C. M. Hoadley, and M. C. Linn, Design-based research in education, in *Internet Environments for Science Education* (AAPT Physics Education, 2004), pp. 73–85.
- [53] S. McKenney and T. C. Reeves, *Conducting educational design research* (Routledge, Abingdon, Oxon, 2013).

- [54] S. Barab and K. Squire, Design-based research: Putting a stake in the ground, *J. Learn. Sci.* **13**, 1 (2004).
- [55] A. A. DiSessa and P. Cobb, Ontological innovation and the role of theory in design experiments, *J. Learn. Sci.* **13**, 77 (2004).
- [56] A. Hofstein and V. N. Lunetta, The role of the laboratory in science teaching: Neglected aspects of research, *Rev. Educ. Res.* **52**, 201 (1982).
- [57] R. Millar, Practical work, in *Good Practice in Science Teaching: What Research Has to Say*, edited by J. Osborne and J. Dillon (Open University Press, Maidenhead, England, 2010), p. 108.
- [58] B. A. Crawford, From inquiry to scientific practices in the science classroom, in *Handbook of Research on Science Education*, edited by N. G. Lederman and S. K. Abell (Routledge, London, 2014), pp. 515–541.
- [59] P. L. Lijnse, “Developmental research” as a way to an empirically based “didactical structure” of science, *Sci. Educ.* **79**, 189 (1995).
- [60] M. Méheut and D. Psillos, Teaching–learning sequences: Aims and tools for science education research, *Int. J. Sci. Educ.* **26**, 515 (2004).
- [61] C. F. J. Pols, What’s inside the pink box? A nature of science activity for teachers and students, *Phys. Educ.* **56**, 045004 (2021).
- [62] M. Araceli Ruiz-Primo and E. M. Furtak, Informal formative assessment and scientific inquiry: Exploring teachers’ practices and student learning, *Educ. Assess.* **11**, 237 (2006).
- [63] W. Ottevanger *et al.*, *Kennisbasis natuurwetenschappen en technologie voor de onderbouw vo: Een richtinggevend leerplankader* (SLO, nationaal expertisecentrum leerplanontwikkeling, 2014).
- [64] W. Spek and M. Rodenboog, *Natuurwetenschappelijke vaardigheden onderbouw havo-vwo* (SLO, nationaal expertisecentrum leerplanontwikkeling, 2011).
- [65] C. F. J. Pols, P. J. J. M. Dekkers, and M. J. de Vries, Introducing argumentation in inquiry—a combination of five exemplary activities, *Phys. Educ.* **54**, 055014 (2019).
- [66] C. F. J. Pols, A teaching sequence on physics inquiry, [10.5281/zenodo.5761998](https://zenodo.org/record/5761998) (2021)
- [67] T. Van der Valk, A. Mooldijk, and J. Wooning, *Guiding for inquiry learning: The falling cones case*, Quality Development in Teacher Education and Training (2004), p. 220.
- [68] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.020170> for details of the teaching sequence and additional data.
- [69] P. Tamir, Practical work in school science: An analysis of current practice, in *Practical Science: The Role and Reality of Practical Work in School Science*, edited by B. E. Woolnough (Open University Press, Milton Keynes, PA, 1991), pp. 13–20.
- [70] E. van den Berg, J. Buning, and T. Smits, *Leren onderzoeken in het voortgezet onderwijs*, *Ned. Tijdschr. Natuurkd.* **271** (1996).
- [71] J. Glaesser, R. Gott, R. Roberts, and B. Cooper, Underlying success in open-ended investigations in science: Using qualitative comparative analysis to identify necessary and sufficient conditions, *Res. Sci. Technol. Educ.* **27**, 5 (2009).
- [72] M. Zion and R. Mendelovici, Moving from structured to open inquiry: Challenges and limits, *Sci. Educ. Int.* **23**, 383 (2012).
- [73] C. Deacon and A. Hajek, Student perceptions of the value of physics laboratories, *Int. J. Sci. Educ.* **33**, 943 (2011).
- [74] A. H. Johnstone and A. Wham, The demands of practical work, *Educ. Chem.* **19**, 71 (1982).
- [75] D. Hodson, Redefining and reorienting practical work in school science, in *Teaching Science* (Routledge, London, 1994), pp. 159–163.
- [76] J. Kortland, Context-based science curricula: Exploring the didactical friction between context and science content, in *Proceedings of ESERA 2007 Conference, Malmö, Sweden*. To be retrieved from the author’s website: <https://www.phys.uu.nl/~kortland/English/Publications> (2007).
- [77] G. Ntombela, A marriage of inconvenience? School science practical work, and the nature of science, in *Practical Work in Science Education: Recent Research Studies*, edited by J. Leach and A. C. Paulsen (Springer, Dordrecht, 1999), pp. 118–133.
- [78] J. K. Gilbert, On the nature of “context” in chemical education, *Int. J. Sci. Educ.* **28**, 957 (2006).
- [79] P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, *Am. J. Phys.* **60**, 637 (1992).
- [80] R. Gott and S. Duggan, Understanding and using scientific evidence: How to critically evaluate data, in *Developing Science and Technology Education*, edited by B. E. Woolnough (Sage Publications Ltd., Buckingham, 2003), p. 146.
- [81] S. Molyneux-Hodgson, R. Sutherland, and A. Butterfield, Is ‘authentic’ appropriate? The use of work contexts in science practical activity, in *Practical Work in Science Education: Recent Research Studies*, edited by J. Leach and A. Paulsen (Kluwer Alphen aan den Rijn, Netherlands, 1999), pp. 160–174.
- [82] B. J. Barron *et al.*, Doing with understanding: Lessons from research on problem- and project-based learning, *J. Learn. Sci.* **7**, 271 (1998).
- [83] R. F. Gunstone and A. B. Champagne, Promoting conceptual change in the laboratory, in *The Student Laboratory and the Science Curriculum*, edited by E. H. Hazel (Routledge, London, 1990), pp. 159–182.
- [84] R. Driver, Constructivist approaches to science teaching, in *Constructivism in Education* (Taylor & Francis Group, Abingdon, England, 1995), pp. 385–400.
- [85] M. Kapur, Productive failure, *Cognit. Instr.* **26**, 379 (2008).
- [86] M. Kapur, A further study of productive failure in mathematical problem solving: Unpacking the design components, *Instr. Sci.* **39**, 561 (2011).
- [87] I. Roll, N. G. Holmes, J. Day, and D. Bonn, Evaluating metacognitive scaffolding in guided invention activities, *Instr. Sci.* **40**, 691 (2012).
- [88] R. Millar, *Analysing Practical Activities to Assess and Improve Effectiveness: The Practical Activity Analysis Inventory (PAAI)*. (Centre for Innovation and Research in Science Education, University of York, York, 2009).

- [89] J. A. Livingston, *Metacognition: An Overview* (2003), <https://eric.ed.gov/?id=ED474273>.
- [90] D. Kuhn and J. Dean David, Metacognition: A bridge between cognitive psychology and educational practice, *Theory Pract.* **43**, 268 (2004).
- [91] M. J. Dehn, *Working Memory and Academic Learning: Assessment and Intervention* (John Wiley & Sons, New York, 2011).
- [92] J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics, *Eur. J. Sci. Educ.* **1**, 191 (1979).
- [93] J. J. S. S. Struble, Using graphic organizers as formative assessment, *Sci. Scope* **30**, 69 (2007).
- [94] C. F. J. Pols, The scientific graphic organizer for lab work, *Phys. Teach.* **62** (2024), [10.1119/5.0094657](https://doi.org/10.1119/5.0094657).
- [95] B. Andersson and F. Bach, Developing new teaching sequences in science: The example of 'Gases and their properties', in *Research in Science Education in Europe* (Routledge, London, 2005), pp. 13–25.
- [96] G. J. Giddings, A. Hofstein, and V. N. Lunetta, Assessment and evaluation in the science laboratory, in *Practical Science*, edited by B. E. Woolnough (Open University Press, Milton Keynes, PA, 1991), pp. 167–178.
- [97] R. Vanderlinde and J. Braak, The gap between educational research and practice: Views of teachers, school leaders, intermediaries and researchers, *Br. Educ. Res. J.* **36**, 299 (2010).
- [98] A. Bryman, *Social Research Methods* (Oxford University Press, New York, 2015).
- [99] H. Altricher *et al.*, *Teachers Investigate Their Work: An Introduction to Action Research across the Professions* (Routledge, Abingdon, Oxon, 2005).
- [100] P. Trowler, *Researching Your Own Institution: Higher Education*, edited by British Educational Research Association Online Resource (2011).
- [101] J. Mason, *Researching your Own Practice: The Discipline of Noticing* (Routledge, London, 2002).
- [102] L. Cohen, L. Manion, and K. Morrison, *Research Methods in Education* (Routledge, London, 2013).
- [103] A. Buffler, S. Allie, and F. Lubben, The development of first year physics students' ideas about measurement in terms of point and set paradigms, *Int. J. Sci. Educ.* **23**, 1137 (2001).
- [104] K. Kok, B. Priemer, W. Musold, and A. Masnick, Students' conclusions from measurement data: The more decimal places, the better?, *Phys. Rev. Phys. Educ. Res.* **15**, 010103 (2019).
- [105] S. Allie, A. Buffler, B. Campbell, and F. Lubben, First-year physics students' perceptions of the quality of experimental measurements, *Int. J. Sci. Educ.* **20**, 447 (1998).
- [106] M. G. Séré, R. Journeaux, and C. Larcher, Learning the statistical analysis of measurement errors, *Int. J. Sci. Educ.* **15**, 427 (1993).
- [107] F. Lubben, B. Campbell, A. Buffler, and S. Allie, Point and set reasoning in practical science measurement by entering university freshmen, *Sci. Educ.* **85**, 311 (2001).
- [108] T. J. M. Smits, *Werken aan kwaliteitsverbetering van leerlingonderzoek: een studie naar de ontwikkeling en het resultaat van een scholing voor docenten*. (Utrecht: CD-β Press, Centrum voor Didactiek van Wiskunde en ..., 2003).
- [109] I. Abrahams, M. J. Reiss, and R. Sharpe, The impact of the 'Getting Practical: Improving Practical Work in Science' continuing professional development programme on teachers' ideas and practice in science practical work, *Res. Sci. Technol. Educ.* **32**, 263 (2014).
- [110] G. Roorda, P. Vos, and M. J. Goedhart, An actor-oriented transfer perspective on high school students' development of the use of procedures to solve problems on rate of change, *Int. J. Sci. Math. Educ.* **13**, 863 (2015).
- [111] R. Boohan, The language of mathematics in science, *Sch. Sci. Rev.* **97**, 15 (2016).
- [112] V. Wong, Variation in graphing practices between mathematics and science: Implications for science teaching, *Sch. Sci. Rev.* **98**, 109 (2017).
- [113] K. W. Kok, *Certain about Uncertainty*, *Mathematisch-Naturwissenschaftlichen Fakultät* (Humboldt-Universität zu Berlin, Berlin, 2022).
- [114] A. Schang, M. Dew, E. M. Stump, N. G. Holmes, and G. Passante, New perspectives on student reasoning about measurement uncertainty: More or better data, *Phys. Rev. Phys. Educ. Res.* **19**, 020105 (2023).