Defining and assessing understandings of evidence with the assessment rubric for physics inquiry: Towards integration of argumentation and inquiry

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 1 July 2021; accepted 15 December 2021; published 15 February 2022)

Physics inquiry can be interpreted as the construction of a cogent argument in which students apply inquiry knowledge and knowledge of physics to the systematic collection of relevant, valid, and reliable data, creating optimal scientific support for a conclusion that answers the research question. In learning how to devise, conduct and evaluate a rigorous physics inquiry, students should learn to choose and apply suitable techniques and adhere to scientific conventions that guarantee the collection of such data. However, they also need to acquire and apply an understanding of how to justify their choices and present an optimally convincing argument in support of their conclusion. In this modified and augmented Delphi study we present a view of inquiry knowledge and a way to assess it that acknowledges both of these components. Using our own expertise with teaching physics inquiry and using curriculum documents on physics inquiry, "inquiry knowledge" is deconstructed as a set of "understandings of evidence" (UOE)—insights and views that an experimental researcher relies on in constructing and evaluating scientific evidence. While insights cannot be observed directly, we argue that their presence can be inferred from a student's actions and decisions in inquiry, inferred with more definitude as a more explicit and adequate justification is provided. This set of UOE is presented and validated as an adequate, coherent, partially overlapping set of learning goals for introductory inquiry learning. We specify conceivable types of actions and decisions expected in inquiry as descriptors of five attainment levels, providing an approach to assessing the presence and application of inquiry knowledge. The resulting construct, the assessment rubric for physics inquiry, is validated in this study. It distinguishes nineteen UOE divided over six phases of inquiry. Preliminary results suggesting a high degree of ecological validity are presented and evaluated. Several directions for future research are proposed.

DOI: 10.1103/PhysRevPhysEducRes.18.010111

I. INTRODUCTION

An important part of physics education at all levels is learning how to *do science*, i.e., to engage in inquiry and develop experimental expertise [1]. In learning how to do science, students engage in *practical work*, small group experiments in which they manipulate instruments and materials to answer a research question [2,3]. In teaching students how to do science and supporting them in developing the required knowledge, it is helpful to see inquiry as the building of a scientifically cogent argument [4–6]. Weighing evidence, assessing alternative methods and explanations of the observed phenomenon, interpreting data, using underlying theories to support the investigative methods and ideas, proactively defending claims against potential criticism by setting limits to the conclusions—all of these actions are components in the construction of a scientific argument [4,7,8]. To be able to produce a scientifically cogent argument that can withstand the scrutiny of (other) scientists, one first needs to understand what it entails to substantiate a scientific claim on the basis of empirical evidence. In this study, we consider a particular kind of inquiry in physics where a quantitative relation between variables is sought. Throughout the paper we refer to this type of inquiry as quantitative physics inquiry (QPI). We first define the understandings required to carry out this type of inquiry. We present these as the learning goals in introductory activities directed at inquiry learning.

Learning goals acquire value only if we are able to measure to what extent students attain them. Objective assessment plays an essential role in enabling students to expand their existing knowledge and ability to gradually plan and devise successive inquiries more effectively [9–12]. However, tools for measuring student's understanding of experimental physics are scarce [13]. Frequently used

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

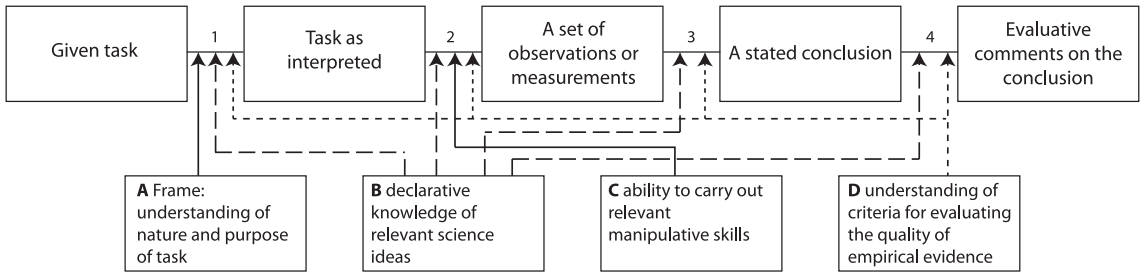

FIG. 1. In their PACKS model, Millar *et al.* [18] link how the taking of decisions during various stages of an inquiry is informed by different types of knowledge.

instruments for assessment in physics lab courses are the Physics Lab Inventory of Critical thinking (PLIC) and the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS), used to determine to what extent students' attitudes and beliefs about physics experimentation concord with those of scientists [9,14,15]. The Scientific Abilities Assessment [16], used to assess whether students can design and conduct a scientific inquiry, is highly regarded for good reason. What this study intends to add to these instruments that evaluate the scientific quality of students' *choices* in designing and conducting QPI, is an evaluation of the presence and quality *of reasons and justifications on which those choices are based*. Assessment tools for physics inquiry other than those mentioned above seem to focus on communication skills such as properly drawing graphs that adhere to scientific conventions, rather than the understanding of, e.g., what makes a particular graph or data representation the most appropriate [1,13,17]. There remains a need for standardized, objective assessment criteria and instruments to assess the degree to which students develop inquiry understandings and skills [9,13]. In this study, we construct an approach to derive student's grasp and use of the proposed understandings from the substantiations and justifications of choices they make during inquiry.

II. THEORETICAL FRAMEWORK

We discuss the role of argumentation in inquiry, specifically in physics, and review what *learning to do science* entails using a theoretical model known as Procedural and Conceptual Knowledge in Science (PACKS) [18]. Subsequently the idea of *understandings of evidence* is introduced to denote the insights, principles and procedures an experimental researcher relies on in constructing, presenting and evaluating scientific evidence for QPI. These are basic understandings we want students to develop.

A. The role of argumentation in learning to do science

Students' physics inquiries have the potential to acquire (scientific) quality only if the students have sufficient content knowledge and apply it appropriately. However, Millar *et al.* [18] argue that for students to effectively engage in doing science, access to appropriate content knowledge is not enough. Students first need to understand the purpose of a scientific inquiry, invest the effort required to produce a scientifically convincing answer to the research question, and understand how to produce trustworthy evidence. In each step of the inquiry the pros and cons of various options are to be recognized and evaluated, and a decision is needed towards attaining optimal cogency within the given constraints (time, money, available equipment, safety). That is, the researcher needs to find a balance between the need to obtain maximum *certainty* about the reliability and validity of the final answers and the limits imposed by *feasibility* of obtaining it. Students should come to understand and feel that from a scientific point of view, inquiry is pointless unless its result is a claim that is as cogent as it can be [19].

This idea highlights the importance of argumentation. Described as the process of reasoning systematically in support of an idea or theory or as "*the uses of evidence to persuade an audience*" [20], argumentation lies at the heart of science and scientific inquiry and thus deserves a central place in science education in general and in scientific inquiry specifically [6,21–25]. While the way students *collect* valid and reliable data is often included in current assessment, how they *substantiate and justify* their choices in establishing these methods is often not (adequately) assessed. So what students do in QPI and how they do it is usually assessed, but apparently further integration of argumentation in inquiry is prevented by a lack of attention for *why* doing so is a good idea, scientifically speaking.

This is what we address in this study. We present, in general terms, the norms and standards against which physicists decide whether a QPI is performed properly, and whether the argument is convincing. We develop a tool for assessment of students' grasp and use of these norms and standards. The building blocks that contribute to constructing, analyzing, judging, criticizing, and improving the cogency of the evidence are recognized in the Procedural and Conceptual Knowledge in Science (PACKS) model as the so-called concepts of evidence [4].

B. PACKS and the concepts of evidence

In their PACKS model presented in Fig. 1, Millar *et al.* [18] distinguish four different types of knowledge (A–D) as

relevant to students conducting inquiry independently. Knowledge type A involves the purpose of the inquiry, e.g., understanding the purpose and nature of the task. Knowledge type B pertains to the relevant content, e.g., understanding the science that is involved. Knowledge type C encompasses the required manipulative skills, e.g., knowing how an instrument should be used. Knowledge type D pertains to the quality of the scientific evidence, e.g., understanding how evidence is derived from data.

In their study, Millar *et al.* [18] conclude that PACKS knowledge type D crucially influences the quality of the inquiry. Important elements of knowledge type D are the so-called concepts of evidence (COE), "concepts that underpin the collection, analysis and interpretation of data" [26–28]. Their tentative list comprises so far 93 concepts such as *fair test ("in which only the independent variable has been allowed to affect the dependent variable")*, range ("*a simple description of the distribution and defines the maximum and minimum values measured")*, trueness or accuracy ("*a measure of the extent to which repeated readings of the same quantity give a mean that is the same as the 'true' mean")*, that underpin the more abstract concepts of the validity and reliability of an inquiry [26]. Gott *et al.* [26] point out that not all of the COE need to be understood in every inquiry. However, some COE play a role in virtually every inquiry, and even the most basic inquiry tends to involve a wide range of COE. According to the authors, these concepts need to be understood before scientific evidence can be handled effectively. It thus seems reasonable to develop and assess students' understanding of each COE and their ability to apply these adequately during an inquiry [4,29–33].

However, there is a complication in assessing students' understanding of each separate COE. Individual concepts acquire their meaning through a network of interrelated concepts [34], so that defining a COE often requires several other COE. It is hard to see how, e.g., one can assess a student's understanding of the concept of dependent variable (39) independently from assessing his understanding of the concepts of independent variable (38) and control variable (47), and perhaps even that of fair testing (46). In developing students' inquiry knowledge, we believe that these COE acquire meaning concurrently and interdependently. Rather than assessing whether isolated COE are present in the student's mind and are applied correctly in an inquiry, we propose to consider groups of COE that are loosely interrelated into meaningful and coherent, partially overlapping wholes that deserve to be called understandings of evidence (UOE). The UOE comprise of the knowledge against which we evaluate the quality of the argument presented in the inquiry, as far as the reliability and validity of the data are concerned. UOE (ought to) guide the actions and decisions of the researcher in constructing that quality. Each UOE may express properties of the evidential information at a particular stage, procedures for constructing that information, as well as prescriptions for enhancing or assessing informational quality.

III. ASSESSING UNDERSTANDING OF EVIDENCE

We present a simple and familiar physics experiment and illustrate how it would usually be assessed. We then ask questions that we believe ought to be answered but generally are not, highlighting what this study intends to add to conventional assessments of inquiry.

> *Consider an inquiry in which a student, age 14, tries to determine how to make a pendulum swing faster. The student uses a 1,0 m long cord and attaches a single mass piece. In her logbook she states: 'A length of 1,0 m makes the movement large enough and the swinging slow enough to allow for suitable time measurements'. She measures how long it takes the pendulum to complete ten full swings using a so-called break beam sensor. After this first measurement, the student increases the mass incrementally by hanging four additional mass pieces from the cord, one by one, performing a single measurement of ten swings each time. It is noted that she hangs the further mass pieces next to, rather than below the first. She is seen to use a protractor, and takes care that each new swing is started from the same angle.*

Conventional assessment would typically focus on this student's mastery of skills: using adequate measuring techniques [using the full available range, (failure to) repeat and average measurements, use of precision instruments, etc.], appropriate handling of data and error (measuring ten swings and dividing by ten so as to minimize measurement error) and maintaining conventions when reporting the results such as using a suitable structure for the report, describing the method in such a way that others can reproduce it, using appropriate graphs [35], and so on. Although these are all relevant aspects of assessment, conventional assessment does not address the following questions that we think are relevant:

1. What made her decide to choose the duration of a full swing (i.e., the *period*) as the relevant quantity to measure "how fast the pendulum swings"?
2. What does she consider to be "large enough" and "slow enough" in the justification of the cord's length, and why?
3. Why did she choose this specific instrument to measure the time of ten swings? Is this choice an optimal choice in light of the research goal?
4. Why did she not repeat each measurement a few times and average the result?
5. Did she have a reason to measure ten full swings instead of one? If so, what reason?

6. Did she have a reason for hanging mass pieces next to rather than below each other? If so, what reason?
7. Why did she measure with 1–5 mass pieces? Is she confident that this suffices to establish the relationship reliably, if it exists? If so, what is that confidence based on?

The student's answers to these questions would inform us about her understanding of how evidence is derived from data [26]. She may have decided to measure the period on the basis of experience rather than understanding and to measure ten swings rather than a single one merely because she has been told to do so in similar situations. However if she can explain that she expected and verified the period to be constant (while the speed of the bob is not), and that ten swings take long enough to minimalize the error due to reaction time, it tells us a lot about her understanding of evidence and the role of data in inquiry. We propose to describe the basic knowledge and understanding that would allow students to answer questions of this kind appropriately.

Whether a student has a specific understanding cannot be directly observed but, we will argue, can often be inferred from her actions and decisions, and can be inferred with more definitude if she justifies those actions and decisions. Note, however, that what is considered an "adequate" attainment level depends, in addition to the expected proficiency level, on the complexity of the inquiry. The expected level of operationalization of each UOE, i.e., what is regarded as needed in producing an optimally convincing answer, depends on the task and the research context—where we assume that the complexity of the task organically grows with the students' age and proficiency level. The following provides an analysis of the kind we propose for this particular inquiry and educational level (with UOE highlighted in **bold**, the COE underlined):

> *The student measured the period with five different masses, controlling the length and starting angle. From this, we infer that she understands that* ***the inquiry is an attempt to establish the relationship (or lack of one) between an independent variable (38) and a dependent variable (39)****. We infer that she is likely to understand as well that when trying to establish such a relation,* ***other variables (****than mass****) might influence the outcomes and should be controlled (47), and that a fair test (46) is needed****. The way she hangs extra masses next to each other, preserving the length of the cord, and uses a protractor reinforces our tentative inference. She measured with an accurate timing device and measured ten swings rather than a single one. We infer that she is likely to understand that* ***it is important to choose suitable instruments and procedures to get valid data with the required accuracy (18) and precision (20).*** *However only a substantiation of her choices would provide certainty on the level of her understanding. While she provides some justification for her choice of length of the cord it would be relevant to know whether her notion of 'suitable' takes into account human error (13), inherent variability of measurements (19) and refers to attaining optimal reliability of the data (14-16). A break beam sensor is a suitable choice of instrument (15) in this experiment, but she may have over-designed it. A simpler instrument, if available, would have sufficed if all she wanted to check is whether the period of the pendulum depends on the mass of the bob.*

This exemplar illustrates some of the understandings students draw on in doing QPI. Some UOE can be inferred from student's actions and decisions: self-initiated systematic variation and control of quantities combined with measurement of another quantity is inconceivable without some understanding of types of variables and fair testing. Other understandings can only be attributed to the student with certainty if she provides more substantiation, but the point is we *want* to be able to assess these understandings, while conventional means do not allow it. Finally, while conventional assessment would register the student's failure to repeat measurements, it would tell us merely that she failed. Merely addressing the symptom by instructing her to "repeat and average" would not suffice. What we would *like* (formative) assessment to accomplish is to point out that an understanding appears to be lacking: there is an inherent variability in measurements in physics, of which the size needs to be established and reported in order to make the answer to the research question trustworthy. We would like to identify this UOE as relevant, establish the level of its attainment and address that if necessary. Many physics teachers will recognize what happens if we do not: students repeat every measurement three times (or five), whether that makes sense or not.

We provided a superficial and incomplete description of a QPI that might occur at the very start of this student's career in science. We might find her at our university a few years later, studying physics and being tasked to determine the acceleration due to gravity within a 0.1% margin of error, by using a pendulum once again. To do so, she would have to operationalize her knowledge at a much higher level, involving a more sophisticated understanding of mechanics, of instruments and measuring procedures, and of the relationship between scientific data and evidence. As regards the latter, this study is meant to describe a set of UOE that is adequate in both situations, and a way of establishing her level of understanding of each UOE irrespective of where she is in her career.

IV. AIMS AND RESEARCH QUESTIONS

In order to assess students' inquiry knowledge we first (need to) define a set of UOE that is necessary and sufficient in devising, conducting, and evaluating basic

TABLE I. The participants in the modified and augmented Delphi study and the research rounds they were involved in.

Participants	No.	Round 1 literature review	Round 2 Delphi Iteration 1	Round 3 field test	Round 4 Delphi Iteration 2	Round 5 expert interviews
Content experts	8		Questionnaire		Interview	Interview
Experts of practice	5			Interview		
External experts	6					Interview

inquiry in physics. We consider the UOE required in QPI: the inquiries that involve the establishment of a relationship between variables. While this includes the vast majority of physics inquiries at secondary school and at introductory physics lab courses, we hope to extend its applicability to other types of inquiry in time. Our first research question therefore is

1. What are the understandings of evidence required to successfully design, conduct, and evaluate physics inquiry in which a quantitative relation between variables is to be determined?

We regard these UOE to be among the learning goals in introductory activities directed at inquiry learning. The second aim in this study is to propose, validate and test an approach to derive the presence and attainment level for each UOE from students' work:

2. What are the characteristics of a valid, reliable, sufficiently specific, and detailed assessment of students' UOE in physics inquiry?

V. METHOD

We first discuss our research design, a modified and augmented Delphi study where we use five rounds to build, review, test, and improve the instrument and subsequently test its ecological validity. We then present for each round the experts, instruments, and analysis involved. Finally, we discuss how we tested the ecological validity of our assessment rubric for physics inquiry (ARPI).

A. Design

The goal of this study is to develop content and construct validity of ARPI where *content validity* refers to the extent that the content covered is indeed the content it purports to cover, and *construct validity* to the extent that the construct measures what it purports to measure [36]. Our approach in this early stage of development is, first, to obtain *direct validity* based on consensus about the theoretical content of the construct between a group of relevant experts [37,38]. Second, to explore *ecological validity* of the construct when it is applied in practice. A reliable and accepted development method in qualitative research aimed at reaching group consensus between experts is the Delphi study [39], an iterative method for the systematic solicitation and collection of judgements by experts on the validity of a construct through a set of carefully designed instruments [40]. Experts' input can be obtained by questionnaires or other means of data collection [41]. In a *modified* Delphi technique experts are presented carefully selected items stemming from, e.g., a literature study [41,42] that eliminates the traditional first round questionnaire, and solidly grounds the study in previously developed work. The modified Delphi approach is likely to reduce the number of iterations required. In subsequent iterations the experts' views are asked and used to adjust, discard, or add items so as ultimately to reach consensus between them [39]. The required number of iterations depends on how quickly experts' views converge. While often three iterations suffice [39,42], sufficient convergence in this study (Table I) was attained in round 4, after two. Iterations 1 and 2 of the modified Delphi section of the study involve the content experts and take place in rounds 2 and 4, respectively. Rounds 3 and 5 explore ecological validity and involve field testing of the construct and expert interviews, respectively, and augment the modified Delphi approach. Rounds 3 and 5 involved, additionally, experts of practice and external experts. Along with the main instruments and experts involved, each round is discussed in detail below. Since the research design includes a modified Delphi study and choosing the appropriate experts is seen as the most important step in this type of design [39], it is convenient to describe the different kinds of expert participants alongside the successive rounds of the design.

B. Participants, instruments, and analysis

1. First round: Prototype on the basis of personal professional expertise and literature review

Goal: The first round aimed at constructing a prototype version of ARPI built on our personal experience with doing and teaching physics inquiry. A supporting literature study ensured that ARPI is grounded in international curricula.

Instruments and procedures: Informed by our personal experience with doing and teaching physics inquiry and by the PACKS model, we produced a tentative list of UOE. So as not to omit relevant learning goals we compared this list with competences and learning goals documented in salient curricula and curriculum related documents, described in detail in the results section. Comparison with available literature to inform the construct's content is in accord with recommendations by McNamara and Macnamara [43] as it

TABLE II. Description of the eight participating content experts in terms of four expert criteria. The symbols *p*, *c*, and *i* denote that the criterion was satisfied in the past, currently or in progress, respectively. Symbols *s* and *u* denote secondary school level and university level, and *d* denotes a Doctorate in physics or in physics education.

Expert	Physics teacher	University Lab course teacher	Physics teacher trainer	Ph.D.
1	*p*, *s*	*c*	*c*	*d*, ed
2	*p*, *s*	*c*		*d*, ph
3	*c*, *u*	*c*		*d*, ph
4	*c*, *u*	*c*		*d*, ph
5	*p*, *s*		*c*	
6	*p*, *s*		*c*	*i*, ed
7	*p*, *s*		*c*	*i*, ed
8	*c*, *s*			

potentially reduces the number of required iterations. We expected the instrument to have acquired *face validity.*

2. Second round: Delphi iteration 1: Acquiring input from content experts

Goal: In order to confirm face validity and to fine-tune the instrument, *content experts* scrutinized the rubric and critically reflected on the relevance, completeness, and clarity of the learning goals and levels of attainment, based on an open-ended questionnaire.

Participants: Content experts need to know what specific knowledge is required to engage meaningfully in QPI. They are required to be experts in teaching and assessing that content. Since this expertise is eminently found among experimental physics researchers and physics educators, our content experts were selected by means of criterion sampling [36]. Eleven physics (lab course) teachers from one network of Dutch secondary school physics teachers and a second national network of university lab course teachers were invited to participate through an email that explained the purpose of the study and the rubric. A representative sample of eight *content experts*, characterized in Table II, agreed to participate. The sample size is well within the range of three to ten recommended by Rubio, Berg-Weger [44].

Instrument and procedure: After they agreed to participate the content experts were sent the rubric, a questionnaire, and an explanatory letter. The letter clarified the aim of the rubric as an instrument to establish students' attainment level of each UOE on the basis of their actions, decisions and justifications regarding QPI. It informed the experts that "UOE" are defined as "the insights, principles and procedures an experimental researcher relies on in constructing, presenting and evaluating scientific evidence."

The content experts were then asked whether they concur with the way the basic understandings of evidence have been described under the heading "The researcher understands that…" and to identify any essential understandings that were missing from the list. These two questions relate to content validity as they deal with the completeness and relevance of the UOE. Furthermore, experts were asked whether they *concur with the specification of the respective observable implications related to the UOE*. They were asked to *consider whether the descriptors per attainment level were clear*, and *whether the three attainment levels were sufficiently distinctive to allow for an objective score*. As these questions address the ability to adequately measure what ought to be measured, i.e., students' attainment levels, they relate to construct validity.

Analysis: Once all data were collected, answers were to be categorized as "consent," "conditional consent," or "dissent." We interpreted the experts' suggestions in terms of the learning goals pertaining to successfully designing, conducting, and evaluating physics inquiry in which a quantitative relation between variables is to be determined. Per suggestion, we analyzed whether multiple experts held the same or contrary views, whether the suggestion was in line with the aims of ARPI and the underpinning ideas, and whether it concurred with relevant literature on physics inquiry and scientific inquiry. We adapted the rubric to improve clarity, completeness, consistency, and applicability, with the ultimate goal of creating consensus on the quality of the content and the applicability of the rubric. Our interpretation of the experts' comments and their view on the adequacy of ARPI as presented in the results section was validated in round four by presenting these interpretations and responses to the same experts and inviting their views.

3. Third round: Exploring ecological validity—Test of ARPI in the field

Goal: This round augments the modified Delphi method that involves the development of content and construct validity, with data on practical applicability, i.e., on the *ecological validity* of ARPI. Furthermore, we considered that gaining insights on how ARPI functions in the field potentially reduces the number of iterations required.

Participants: Twenty teaching assistants (TAs) participated in a training session directed at identifying problems with application of ARPI, suggesting and discussing potential solutions to these problems, and implementation of these potential solutions in an authentic setting. Five of the TAs were subsequently interviewed to evaluate the effectiveness of the attempted solutions and to identify remaining issues. These five TAs are seniors, in their third year or higher, and are considered to be *experts of practice* (Table II) in terms of the practicality and application of ARPI. They supervised less senior TAs and were therefore aware of actual and potential problems generally encountered by TAs in assessing lab reports.

Instrument and procedure: To test the applicability of the instrument and to establish the conditions that make it

applicable in practice, the revised version of ARPI was applied in the introductory physics lab course at our university. To let the TAs get acquainted with ARPI's content, and to train them in using this new assessment form, a training session was conducted. All 20 TAs of the course graded a sample report as part of their training. The problems they encountered in objectively grading the sample report were identified during a subsequent evaluative session. We proposed solutions to these problems including adjustment of the rubric and extra training, and implemented these if the TAs considered them promising. The TAs then applied ARPI in the regular course by grading the lab reports of 70 students. Finally, a *semi-structured interview* was used to obtain the senior TAs' views on the applicability of the revised version of ARPI. The interview focused on two questions: how did they and those they supervised experience assessment with ARPI and did they (still) encounter problems when ARPI was applied in the regular course after the training exercise.

Analysis: Remaining problems with grading were identified as potential "threats" to ARPI's adoption in actual educational settings. From the identified problems and effective solutions were inferred the conditions that ought to be met for the instrument to become optimally applicable.

4. Fourth round: Delphi iteration 2—Determining the consensus between content experts

Goal: The final revised version of ARPI was once more inspected by the content experts in order to establish its content and construct validity. In interviews, changes were discussed, remaining and emerging issues were addressed, and consensus was sought or confirmed.

Instrument and procedure: Content experts were provided with the revised rubric, with all modifications highlighted so as to present the group responses. They inspected it well ahead of the interview. The interview protocol guided the discussion of, first, the general modifications. Experts were asked whether they accepted these. Next, the experts reflected on their own previous answers. Their specific round two comments were read and our interpretations and responses (e.g., a modification of the rubric) presented. Where necessary, the purposes of ARPI were revisited, and our response provided with a rationale or justification. Experts were given the opportunity to discuss whether they perceived their previous input to be adequately and sufficiently dealt with. They were invited to discuss whether they had identified new issues of concern, and to forward any essential additional understandings they believed ought to be included.

Analysis: The experts' answers were again categorized as "consent," "conditional consent" or "dissent." These data, to be found in the results section, allowed us to establish the level of consensus about the rubric as a specification of learning aims of physics inquiry and about its function as an instrument to measure the attainment levels of these aims.

We consider to have achieved consensus on content and construct validity if at least 80% of the experts concurred with the final version of ARPI. This is in accord with the criteria of defined consensus as elaborated by Miller [45] so that no further iteration in the modified Delphi part of the study was deemed necessary. Remaining contentious issues are presented so as to illustrate potential areas of further development.

5. Fifth round: Exploring ecological validity: Expert interviews

Augmenting the field test in terms of ecological validity, we explored whether the *content experts* regarded ARPI to have added value with respect to conventional inquiry assessment methods. Semistructured, live interviews based on two open-ended questions were conducted to explore whether they would consider using ARPI in their own educational practice, and what reasons they had for either considering it or not. The same questions were put to a third team of six external experts (see Table I). This group of PhDs in physics were found by means of convenience sampling from a faculty online learning community (FOLC) [46]. Their involvement contributes to the external validity of ARPI [47,48] as five of them are principal lecturers in one or more upper-level university physics lab courses at universities across the USA. We looked for emergent themes in the answers based on content analysis (Cohen, Manion, and Morrison [36], pp. 674–685).

6. Ethical statement

All experts participated on a voluntary basis on condition of anonymity. They allowed all of their input, including input provided in video recorded interviews, to be used for research purposes.

VI. RESULTS

The results obtained in the five rounds are presented consecutively. We first highlight the main features of the prototype version and the literature it is based on. Subsequently, the input provided by the content experts in round two is presented, and then the input from experts of practice in the field test of round three. Content and construct validity based on content expert consensus about the final version of ARPI are discussed as the main outcome of this study in round four. Finally, we present ARPI and data pertaining to its ecological and external validity derived from round 5.

A. First round: Prototype on the basis of personal professional expertise and literature review

A first tentative list, constructed from our personal professional knowledge and experience in doing and

TABLE III. The phases of ARPI overlap with the phases as distinguished in the APU model and by Kempa.

Kempa	APU	ARPI
Recognition and formulation of the problem	Problem formulation	Asking questions
Design and planning of experimental procedure	Planning an experiment	Design
Setting up and execution of experimental work (manipulation)	Carrying out an experiment	Methods and procedure
Observational and measuring skills (including the recording of data and observations)	Recording data	
Interpretation and evaluation of experimental data and observations	Interpreting data and drawing conclusions	Analysis
	Evaluation of results	Conclusion and evaluation

teaching physics inquiry, adapted to the PACKS framework, consisted of 16 UOE. To structure ARPI, we divided the UOE over the various phases of inquiry. To do so, we found a convenient structure by considering the phases distinguished in the assessment of performance unit (APU) model on which the PACKS framework is built [18,49,50], and Kempa's model of doing science, which is recognized to be useful for assessment [11,38]. As shown in Table III, the phases of ARPI integrate the phases of the other two models.

We constructed UOE by considering loosely interrelated groupings of COE and attempting to identify the main themes or principles present in these relations. In formulations of the UOE the COE are sometimes explicitly recognizable (see the two examples in the analysis of the exemplar), but not always. For example, UOE 12 reads *The researcher understands that data require appropriate methods for analyzing and describing them*. These "appropriate methods" subsume COE 62-82, including tables, bar charts, the linear regression method, line of best fit, etc., without enumeration.

The construction of ARPI distinguishes carefully between the UOE present in the researcher's mind and the actions guided by these UOE. The column headed "*The researcher understands that …*" refers to the UOE while the column detailing the actions, decisions and justification informed by these UOE is headed "*This understanding is demonstrated by …*." The first tentative list of familiar learning goals and aspects of inquiry learning was rendered more authority by comparing it with the literature on physics curricula and curricular recommendations for the secondary and tertiary level.

Compulsory secondary physics education mainly aims at developing scientific literacy [32,51–55]. The *Program for International Student Assessment* (PISA) is geared towards assessing scientific literacy internationally. The basis for its 2015 implementation is the 2015 Draft Science Framework presented by the *Organisation for Economic Co-operation and Development* (OECD) (2013). Two of the framework's three core abilities of scientific literacy relate to inquiry: Evaluate and design scientific inquiry and Interpret data and evidence scientifically. In presenting the Next Generation Science Standard (NGSS), a Framework for K–12 science education, the National Research Council (NRC,2013) specifies eight essential practices of science and engineering. Dutch curricula for secondary science, in particular physics, are heavily influenced by, show similarities with, or paraphrase these two documents [56–58]. Other international curricula in the English-speaking world are similarly derived from these sources [59–63]. Therefore, we consider the OECD [32] and NRC [52] documents to be adequate and sufficient in their description of the learning goals for secondary school level physics inquiry.

At the tertiary level, physics education aims at teaching students to *think like a physicist* [64–66]. Wieman [67], Nobel laureate in physics, provides a list of cognitive activities that a physicist goes through during experimental research. A more detailed list of learning outcomes related to the undergraduate physics laboratory curriculum is provided by the American Association of Physics Teachers (AAPT) Committee on Laboratories [65]. Furthermore, the frequently referenced source Etkina *et al.* [16] defines *scientific process abilities* for introductory physics students. We consider the combination of these documents to provide a representative set of learning goals for tertiary physics inquiry.

Any learning goal in these five documents relevant to successfully designing, conducting and evaluating physics inquiry was included in ARPI if found to be absent and yet related to the reliability and validity of data. As an example of the process consider Table IV[52]. Most NGSS goals matched the UOE of the prototype list, but "*planning and conducting an investigation in a safe and ethical manner*," although highly important, was not adopted in the list as it relates to aspects and understandings other than those of the reliability and validity of the evidence which ARPI is meant to assess.

On the other hand, the initial list did not include the provision and reception of feedback although, as Driver, Newton, and Osborne [5] state "*It is through such processes of having claims checked and criticized that 'quality control' in science is maintained*." We therefore included an UOE specifying that scientific knowledge is a product of intensive consultation and discussion between experts judging the evidence for the stated claim. Utilising (peer) feedback is a powerful instrument in improving the quality

TABLE IV. Comparing the UOE with the learning goals found in the NGSS revealed that receiving and providing feedback was missing.

Practice 3: Planning and carrying out investigations	UOE
Plan an investigation or test a design individually and collaboratively to produce data to serve as the basis for evidence as part of building and revising models, supporting explanations for phenomena, or testing solutions to problems. Consider possible confounding variables or effects and evaluate the investigation's design to ensure variables are controlled.	4–6
Plan and conduct an investigation individually and collaboratively to produce data to serve as the basis for evidence, and in the design decide on types, how much, and accuracy of data needed to produce reliable measurements and consider limitations on the precision of the data (e.g., number of trials, cost, risk, time), and refine the design accordingly.	4–10
Plan and conduct an investigation or test a design solution in a safe and ethical manner, including considerations of environmental, social, and personal impacts.	Not included
Select appropriate tools to collect, record, analyze, and evaluate data.	5, 12–13
Make directional hypotheses that specify what happens to a dependent variable when an independent variable is manipulated.	3

of inquiry. This understanding can be used to improve one's own work as well as to point out weaknesses in the work of others and help to improve it. To acknowledge both aspects of the understanding, this UOE (19 in final version of ARPI) has two aspects: providing feedback, and soliciting and dealing with feedback. To emphasize that this understanding relates to *all* phases of inquiry, we added a sixth phase named "review." No other learning goals in these sources needed to be included.

Assessment of aims of learning requires not only their specification but also the description of attainment levels, while curriculum documents often specify only the highest of these. To establish how many levels were required we consulted the appropriate literature [68–70] but found no consensus [68]. Moskal [69] and Ref. [70] suggest that one can start with a limited but meaningful number of attainment levels and add more later on, if required. We decided on three attainment levels to start with.

We then constructed descriptors for these levels. At the lowest level, the understanding is apparently absent as the actions and decisions are seen as inadequate. At intermediate level the understanding is apparently applied, the actions and decisions are (partly) valid, but are not or insufficiently substantiated. At this level, the actions of a student do not or not fully warrant attribution of the UOE concerned. At the highest level the understanding is adequately applied and substantiated *and* the UOE attributable because the actions and decisions cannot be understood without it.

In the first round we produced a prototype containing 17 UOE as aims of inquiry learning divided over six phases of inquiry with descriptors for three attainment levels within each UOE.

B. Second round: Delphi iteration 1—Acquiring input from content experts

Guided by open-ended questions, the experts were asked to scrutinize the prototype version. Seven experts conditionally accepted our set of UOE as a complete set of inquiry knowledge required to successfully design, conduct, and evaluate QPI. One expert fully concurred. The following is an illustrative example of an expert's reply. He sees the UOE as relevant, but holds that some aspects of understandings remain implicit or ought to receive more attention (translated and paraphrased):

> *I can agree with [the instrument] but am quite attached to terms like 'finding information' and 'communication'. The former I don't find explicitly anywhere (while I think it is indispensable at any level). 'Communication' I recognise only in the final [UOE], while that actually is more concerned with 'feedback'.*

Only one expert raised no issues, all others raised one or more. However only one issue, *assessment of communication*, was raised by two, and none by more than two experts. Table V presents all issues raised, our response, and our rationale for that response. Responses and rationales were presented to the experts in the fourth round, and their reaction is reported there.

There was no agreement between experts on the number of attainment levels. The required or desired number of levels varied between 2 and 5. Because of a lack of consensus among the experts on this issue, resolving the matter was deferred to the next stage of the study.

C. Third round: Exploring ecological validity—Test of ARPI in the field

A training session was conducted for TAs to practice applying ARPI in assessment of inquiry reports. Based on their assessment of a sample report the problems they encountered were identified. One of their problems involved the number of attainment levels for each UOE. The students were found to occasionally outperform one level but not fully attain the next higher level. TAs

TABLE V. Issues raised by the experts in the second round along with our response.

Issue raised	Response	Rationale
Do all inquiries necessarily start with a research question?	Clarification of content.	Interpret "starts with" [71] as "is based on," or "is founded in."
Is asking questions relevant in the given educational settings?	Clarification of aims.	Activities that include "posing questions" have to be assessable by the instrument [1].
I would explicitly include the word "hypothesis."	Adapted by distinguishing UoE3 from UoE1.	If feasible, expectations regarding an experiment indeed ought to be formulated as a hypothesis.
I miss the assessment by means of the lab journal.	Clarification of aims.	Lab journals are not excluded, rather ARPI is meant to assess the lab journal as one source of information on a student's attainment levels.
I miss assessment related to presentation and communication.	Clarification of content & aims.	Issues pertaining to presentation and communication are assessed, but as integral parts of the expression of UOEs.
I miss information related to gathering theoretical information.	Clarification of aims.	Assessing content is not the purpose of ARPI, it is meant only to assess type D knowledge in the PACKS model.
I miss that "unexpected" observations could trigger new inquiries.	Adapted by including UOE 18.	ARPI ought to include understandings pertaining to awareness of needs and options for further research.
I would suggest to include that parameter values should be chosen wisely so as to optimize measurable effects.	Clarification of content.	The choice of appropriate parameters is meant to be understood as part of UOE 6.
I would suggest to rephrase …	Rephrased when appropriate.	Minor rephrasing increases the clarity and consistency of text.

questioned whether allocating scores in between levels was allowed. Combining their remarks with input from the *content experts* it was decided to identify two additional attainment levels.

A second issue that was brought up in the evaluation session involved some TAs expressing a lack of confidence in assigning attainment levels based on their interpretation of the adequacy of the student researchers' decisions. As this insecurity appeared primarily among the more junior TAs and seemed to stem mainly from inexperience in grading and a limited inquiry knowledge, junior TAs were subsequently matched with senior counterparts. They graded inquiry reports as teams so as to discuss and resolve contentious interpretations.

After addressing the two main issues as described above, the next step in exploring the applicability of ARPI in the field involved the grading of 70 first-year physics inquiries. The experts of practice, i.e., the senior TAs, were then asked for feedback in an interview session. The general content of these interviews is adequately summarized by one of them:

> *As an assessor, it takes more time to assess using ARPI because the criteria are less absolute and thus one needs to provide a further substantiation. ARPI also requires a deeper understanding of the inquiry process before one is able to assess the work of others. Although this should not be a problem, it might require some attention.*

The number of attainment levels was no longer an issue for any of the experts of practice. Rather, they felt the approach supported them in providing targeted feedback. The experts regarded ARPI as useful since it focuses on the students' thinking in devising and conducting a physics inquiry, which some saw as a neglected aspect in our traditional assessment:

> *The current form of assessment for physics inquiries lacks various features when [I'm] providing not only a grade but also feedback to a student. However, ARPI aims to fill several of its gaps. It analyses the critical thinking of a student when designing the experiment and analysing the data, where limitations of the experiment are key to determine the validity of its outcome. This allows for feedback which informs the student about his/her stage in becoming a researcher.*

D. Fourth round: Delphi iteration 2—Determining the consensus between content experts

To obtain the *content experts*' view on the revised version of ARPI and discuss remaining and emerging issues, the content experts were interviewed. All experts agreed that, given the findings in the test and our explanations, the use of five attainment levels is justified. According to one expert,

Choosing five levels allows students to proceed from one level to another more easily. It might help students to see their own progression.

Furthermore, all experts agreed that including UOE 18 is sensible and in line with the other UOE. The experts agreed that their specific, individual issues were addressed sufficiently or a proper rationale was provided. The following vignette (paraphrased and translated by the author and approved by the expert) illustrates the discussions in which consensus was sought:

Researcher: You stated that hypothesis testing was missing. We included the word hypothesis in one of the UOE. Given the elaboration of the purpose of ARPI, do you think the issue is still relevant?
Expert: Given the specific aim of establishing the relation between two variables, the issue is not relevant anymore.
Researcher: A second issue you raised is whether an inquiry starts with a research question. I would like to refer to the VASI instrument of Lederman where this view is advocated and this specific sentence is used.
Expert: I guess that whether it actually begins with a research question is a matter of definition, but I think it is justified to use the wording of the literature.

Some new issues were raised that could be dealt with directly. An example,

Expert: None of the UOE seems to relate to student's plan of approach to analyze the data.
Researcher: I think that is covered in UoE4, "the research question should be answerable with the devised experiment," demonstrated by "explaining how planning, collection, evaluation of data relate to the aim of the experiment."
Expert reads the UOE.
Expert: Yes, it is covered in that specific UOE. However, if students are able to explain how they will analyze the data to answer the research question, this would significantly improve other aspects of student's inquiry. You could think of breaking up the UOE in two parts. However, it is just a suggestion.

This expert initiated the discussion that was mentioned in relation to Table V, on whether choosing optimal parameter values should be included. He now noted,

It might be too specific and depends on what kind of experiments you are doing. It doesn't cover all possible kinds of experiment.

The issue was further addressed by inspecting UOE 6. The expert agreed that it largely covers the issue, and considered the issue resolved.

The final construct, presented in Table VI, consists of 19 UOE divided over 6 phases of inquiry. The UOE form a summary of the inquiry knowledge required to successfully design, conduct, and evaluate QPI. Per UOE, five attainment levels are distinguished, where descriptors for the lowest, intermediate, and highest level are worked out in detail. In the fourth round, all content experts accepted the adjustments and approved the rationales we provided to address their specific issues. No new issues other than those discussed above were raised. The descriptors are regarded to be sufficiently clear and distinctive for scoring student's attainment levels.

Since we specified the benchmark for consensus on content and construct validity to be at a minimum of 80% of the experts concurring, we take it that consensus on the final version of ARPI has been established and that the rubric has acquired both content and construct validity.

E. Fifth round: Exploring ecological validity—Expert interviews

Even if the content and construct validity of ARPI are approved, it will not be adopted in actual educational setting unless the educators involved regard that as feasible and worthwhile. The instrument requires *ecological validity* in order to attain its purposes. Therefore representatives of these educators, i.e., the content experts and external experts, were interviewed to establish whether they would consider using the rubric in their practices, and what reasons they would have for either doing so or not. All experts stated that they would like to use (parts of) ARPI and indicated that the rubric adds value to their current assessment methods. To adopt it in their own educational setting various experts suggested, it could be adapted to suit experiments with specific educational purposes and be merged with their current assessment formats in which other PACKS knowledge types are assessed as well. Some secondary school teachers suggested to use ARPI as a learning tool. To facilitate younger students' understanding of all elements in the rubric, they advised rephrasing some of the UOE for that purpose.

The experts offered various reasons for applying ARPI in their own educational setting:

- To grade students who engage in (open) inquiry.
- To augment their current assessment format by, i.e., including elements that are as yet missing and reformulating attainment levels similar to ARPI with a focus on argumentation.
- To review current experiments and specify the learning goals using ARPI.
- To use it as a source of inspiration in designing practicals addressing specific UOE.
- To help students develop their inquiry and use ARPI in a formative way.

TABLE VI. ARPI consists of 19 understandings of evidence applied by a researcher when conducting a physics inquiry. Indicators for the lowest, intermediate and highest level are provided. Levels in between these are assigned when a student outperforms the lower level but has not fully attained the higher level. Table is available in excel [72].

UoE	Phase	The researcher understands that:	This understanding is demonstrated by:	Highest level	Intermediate level	Lowest level
1	1 Asking questions	A scientific inquiry starts with a research question.	Posing a research question that is clear, unambiguous, sufficiently specified and researchable.	Formulates the research question in such a way that it is accessible through scientific research.	Formulates the research question but with a lack of relevant information.	Does not formulate the research question (well).
2		The inquiry is an attempt to establish the relationship (or lack of one) between an independent variable and a dependent variable.	Expressing the research question in terms of appropriate, measurable variables.	Identifies the dependent and independent variables and expresses the research questions in terms of these.	Identifies the relevant variables but fails to relate the experiment to them.	Fails to identify (in)dependent variables or to regard the experiment as a way to determine the relation between them.
3		Expected outcomes are formulated, when appropriate in the form of a testable hypothesis.	Formulating expectations regarding the findings in a substantiated and empirically verifiable form.	Formulates substantiated and testable expectations.	Formulates expectations in a testable but insufficiently substantiated form or in a substantiated but not well testable form.	Does not formulate or substantiate expectations even though these are required or desirable.
		The researcher understands that:	**This understanding is demonstrated by:**	**Highest level**	**Intermediate level**	**Lowest level**
4	2 Design	The research question should be answerable with the devised experiment.	Explaining how planning, collection, evaluation of data relate to the aim of the experiment.	Explains explicitly and in detail how the collection and interpretation of the data will be used to answer the research question.	Accounts for how the data will be used to answer the research question but with lack in detail and/or specification.	Fails to explain (independently) how the experiment allows one to answer the research question.
5		Other variables can affect the dependent one, therefore a fair test is needed, keeping these variables constant.	Identifying relevant variables and controlling them in constructing a fair test.	Substantiates which variables are relevant and how these are controlled in order to use fair testing.	Identifies and controls some but not all of the relevant variables.	Fails to identify or control relevant variables that may affect the dependent variable.
6		It is important to choose suitable instruments and procedures to get valid data with the required accuracy and precision.	Choosing appropriate measuring instruments and procedures that provide the required reliability and accuracy of the dataset.	Makes an informed, substantiated and acceptable choice between instruments and procedures so as to ensure optimally reliable and accurate data.	Considers options regarding instruments and procedures but fails to reach (independently) an optimal choice.	Ignores options for selecting measuring instruments or procedures that would enhance data quality.
7		(Human) Errors and uncertainties may occur and precautions are needed to minimize or avoid them, ensuring reliability.	Identifying sources of uncertainty and error, and taking and justifying precautions.	Takes all relevant causes of uncertainty and error into account and develops or augments procedures to minimize them.	Takes precautions to minimize effects of some but not all sources of uncertainty or error or fails to practically implement the precautions.	Fails to identify sources of uncertainty and error.
		The researcher understands that:	**This understanding is demonstrated by:**	**Highest level**	**Intermediate level**	**Lowest level**
8	3 Method & Procedure	Measured values will show inherent variation and the reliability of data must be optimised, 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.	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.	Repeats measurements a fixed but sufficient number of times without substantiation in terms of the quality of the dataset. Considers collecting additional data only in retrospect, as a recommendation.	Collects too few repeated measurements without substantiation or consideration of the quality of the dataset. Does not consider collecting further data at any stage.
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 an appropriate and sensible measurement range and interval.	Chooses and substantiates appropriate measured minimum, maximum and interval.	Measured minimum, maximum and/or interval are appropriate but lack substantiation.	Measures inappropriate minimum, maximum and/or in-between values.
10		It is important to use instruments and carry out procedures properly to obtain valid data with the required accuracy and precision.	Intentionally carrying out measuring procedures and using instruments appropriately to optimally reduce measurement uncertainty.	Manipulates equipment and instruments purposefully, correctly and systematically in optimizing repeatability and minimizing potential error.	Manipulates equipment and instruments purposefully, correctly and systematically but fails to do so fully continuously and consistently.	Fails to manipulate equipment and instruments purposefully, correctly and systematically.
		The researcher understands that:	**This understanding is demonstrated by:**	**Highest level**	**Intermediate level**	**Lowest level**
11	4 Analysis	In a series of measurements outliers may occur and should be examined and discarded if there is sufficient reason to do so.	In a series of repeated measurements or an observed trend in the data, identifying and dealing with outliers in an appropriate, justified way.	Takes outliers into account, excludes these if appropriate and substantiates this choice. Collects additional data to replace removed outliers if that is feasible.	Excludes outliers when that is sensible but does not add measurements if that is feasible, or does not substantiate exclusion.	Does not consider outliers, treats the measured values as ordinary.
12		Data require appropriate methods for analysing and describing them.	Choosing data representation methods that reveal clearly and unambiguously the properties of, and patterns (or absence of these) in the data set.	Makes use of appropriate data representations, clearly revealing the pattern and features in the data.	Chooses suitable but not optimal data representations to establish a pattern.	Chooses inappropriate data representations.
13		An optimally informative answer to the research question requires a description of relationships in as much detail as possible. Quantitative descriptions are more detailed than qualitative ones.	Describing the data by identifying salient and relevant patterns in detail and if possible their mathematical expression.	Describes patterns in appropriate detail. Specifies a mathematical expression or describes the quantitative relationship of the dataset if possible.	Describes patterns correctly but misses some details of features or mathematical properties in relationships.	Expresses relationships in a qualitative sense only.

(Table continued)

TABLE VI. *(Continued)*

		The researcher understands that:	This understanding is demonstrated by:	Highest level	Intermediate level	Lowest level
14	5 Conclusion and evaluation	A complete, clear, substantiated and useful answer to the research question must be formulated.	Formulating a clear, substantiated and unambiguous answer.	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.	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 an unclear and unsubstantiated answer which is insufficiently informative or insufficiently supported by the data.
15		The reliability of the dataset is to be accounted for by considering how well each datum was measured and the reliability of the established relationship.	Discussing how the design ensures optimal trustworthiness of the data and the outcomes in the given circumstances and specifying limitations to the method, procedures and/or equipment.	Specifies and justifies the quality of the data and the conclusion in terms of how well the data match the relationship that was found and discusses limitations due to the method and/or equipment.	Specifies and justifies the quality of the data and of the conclusion in terms of how well the data match the relationship that was found, but not fully or not adequately.	Does not justify the quality of the data and of the conclusion in terms of how well the data match the relationship that was found.
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.	Adequately substantiates limitations to the validity of the conclusion.	Discusses features and limitations to substantiate the validity of the inquiry and its outcomes, but inadequately or only partially.	Does not discuss features and limitations that address the validity of the inquiry.
17		The quality of the inquiry can virtually always be improved with the gained insights.	Proposing recommendations following from the conclusions of the inquiry with appropriate and explicit emphasis on the most critical limitations.	Provides substantiated recommendations which are shown to address the most important limitations of the inquiry.	Provides recommendations that address important limitations but are not or only partly substantiated.	Provides no relevant, substantiated recommendations.
18		New questions may arise related to the inquiry.	Proposes follow up studies that stem from the outcomes of the inquiry.	Proposes and substantiates relevant follow up studies that build on the outcomes of the inquiry.	Proposes follow up studies that do not constructively or directly build on previous inquiry's findings.	Does not propose any follow up studies.
		The researcher understands that:	**This understanding is demonstrated by:**	**Highest level**	**Intermediate level**	**Lowest level**
19	6 (Peer) Review	Scientific knowledge is a product of intensive consultation and discussion between experts judging the evidence for the stated claim. Utilising (peer) feedback is a powerful instrument in improving the quality of inquiry.	Providing critiques on scientific arguments by probing reasoning and evidence and challenging ideas and conclusions.	Provides constructive feedback, challenges conclusions where possible.	Misses essential points or methods for providing feedback.	Does not provide effective or relevant feedback.
			Solicits feedback, responds constructively and processes effectively critiques of the quality of the scientific argument in improving the inquiry.	Solicits, accepts, and uses feedback to improve the inquiry, or defends it by presenting counter arguments.	Hardly solicits feedback. Some essential parts of the feedback are ignored or not successfully acted upon.	Does not solicit, accept, or use feedback as a way to improve the inquiry.

One external expert, a member of the *AAPT Committee on Laboratories* providing recommendations for the undergraduate physics laboratory curriculum [65], reflected

> *It would help me in designing experiments, where one particular aspect of the rubric can be applied, like treating the aspect of outliers. It makes clear that a specific experiment is targeting a specific aspect.*

In ensuing discussions, several educators questioned whether all items should be assessed in each inquiry and whether ARPI is or could be relevant to other types of (physics) inquiry. Just as with other aims of learning we surmise that ARPI can be used as the starting point for the development of learning pathways in which the aims are approached iteratively by students. Further research will have to show whether a natural order of UOE suggests itself, or a more integrated approach is more effective. It is unlikely that a learning process is effective if it addresses all aims at once, or if it provides no structure and focus, but the details are not known at present. Constructive alignment [73] is indispensable and we hold that ARPI, or the underlying ideas on which the construct is based, is functional in maintaining it.

VII. CONCLUSION

We constructed 19 understandings of evidence which are understood as the inquiry knowledge a researcher relies on in producing, evaluating, and presenting a rigorous physics inquiry in which the relation between two variables is to be determined. We regard these UOE as the learning goals for activities that are meant to develop student's physics inquiry knowledge. In ARPI five attainment levels are distinguished. The highest attainment level is assigned when the student is able to adequately justify and substantiate particular decisions pertaining to the UOE. "Adequate justification and substantiation" were defined in terms of whether the inquiry results in a claim that is optimally cogent from a scientific perspective, in answer to the research question. Intermediate and low levels of attainment have also been specified in terms of conceivable actions, decisions, and justification reflecting each of these levels. The next-to lowest and highest levels did not to require full specification, as determined in field testing. They are assigned when a student outperforms the lower level but does not quite attain the next higher level. A modified and augmented Delphi study was used to acquire content and construct validity of the resulting construct: the

Assessment Rubric for Physics Inquiry. ARPI enables one to assess student's attainment level of physics inquiry, where the focus on student's substantiation of choices emphasizes the central place argumentation plays and deserves in scientific inquiry. ARPI involves assessment of aspects of inquiry that previously were not (fully) considered, and its implementation hence requires training of the assessors. To assign students' attainment levels as objectively as possible, three conditions need to be met: (i) an appropriate attainment level of the assessor, (ii) access to the relevant information (report, lab journal, discussion with students), and (iii) enough time to perform the assessment. Provided these issues are addressed, the preliminary results suggest that ARPI has a high degree of ecological validity as it is considered by the experts to be both feasible and of added value in the relevant educational settings.

VIII. DISCUSSION

This study has both an educational and theoretical yield. It is not difficult to envision the educational value of the validated assessment format that extends current assessment by revealing some of a student's thinking behind the doing [26] and examining whether the decisions and actions are based on inquiry knowledge. Doing inquiry is hard to teach and learn since there is no scientific method that dictates how scientific quality is to be attained. There are methods of science based on insights attained and conventions agreed on by researchers in their field of expertise, and rules meant to facilitate adherence to the conventions and insights. However, while the conventions have been well specified, these insights tend to remain implicit. As a consequence, each new inquiry may be experienced by students as a completely new task in which they have to "discover" why these rules apply. As Millar [74] argues, however, *it may be more feasible to teach students how to evaluate their data and present justifications to support conclusions, than to teach them how to tackle new tasks*. He refers here to the development of students' understanding of PACKS knowledge of type **D** in which the COE are important elements. However, these COE do not acquire meaning one by one but as integrated, preferably meaningful wholes. Meaningful in that students understand *why* these COE matter. Our framework of UOE is meant to enhance knowledge of type **D** by making these coherent, integrated, meaningful understandings explicit. They are the yardsticks scientists use in comparing the quality of decisions and justifications in inquiry: better decisions produce answers to research questions that are scientifically more cogent. ARPI and the associated UOE provide a framework for considering what counts as quality research. The framework is a starting point for building a pedagogical theory in that it describes what understandings students essentially need to develop in creating evidence from observations, and points out how their level of understanding can be assessed on the basis of their actions, decisions and justifications. The premise of this theory is the notion that an inquiry comes down to the building of a scientifically cogent argument where each decision and action undertaken is substantiated. Developing a pedagogical theory of this kind targets the design and implementation of educational activities that progressively develop students' understanding of the criteria to evaluate the quality of empirical evidence [18] on the basis of the understandings specified in ARPI.

IX. LIMITATIONS AND FUTURE RESEARCH

ARPI was constructed with a focus on knowledge type **D** in the PACKS model [18] by organizing interrelated COE [26] into coherent UOE. As is often done in curriculum documents, we considered element of type **D** knowledge in isolation. As the construct relies (almost) solely on type **D** knowledge, it is possible to use ARPI for various kinds of physics inquiries that do not explicitly involve or focus on physics content or in inquiries where the students command the physics content involved. However, in real physics inquiries different types of knowledge are often applied in an integrated way where they interfere with each other [9]. In our field test we successfully applied ARPI without interference of PACKS type **B** knowledge. However further study is required to explore how ARPI can be combined with other assessment formats that focus on PACKS type **B** knowledge in more 'authentic' inquiries. It is worthwhile to investigate how ARPI and its framework can be integrated in models for inquiry—such as the Modelling Framework for Experimental Physics [75,76]—that focus especially on PACKS type **B** knowledge.

The construction and validation of ARPI was restricted to QPI where every UOE was intended to be applicable regardless of the student's level. Further development of the instrument encompassing other types of physics inquiry and other natural sciences is not difficult to envisage but requires further work. In this paper we briefly elaborated its applicability in our first year physics lab course only. A forthcoming paper will present a teaching sequence which aims at the development of key UOE in 14–15-year-old students. Furthermore, ARPI and the UOE are considered for use and further development in the various lab courses throughout the physics program at our University.

While content and construct validity of ARPI have been established qualitatively, its reliability—the consistency or concordance with which a score is assigned—has not yet been quantitatively determined. It is our intent to explore and compare the interrater reliability of untrained and trained TAs in a joint study of two universities, thereby further exploring the conditions that need to be satisfied to use ARPI as an assessment tool. Furthermore, we intend to explore how to equip secondary physics teachers to use ARPI. We are developing a rubric, augmented with

examples, that is formulated in terms also the youngest students can understand, thereby heeding the request of some of the experts to expand the use of ARPI as an assessment instrument to include instructional purposes. We would like to think that ARPI can then help them, or the hypothetical student from our exemplar, to become researchers who understand that they need to substantiate their decisions, explicate constraints, and elaborate on the inquiries' validity and limitations. In other words, that they use argumentation to improve and defend their work, understanding that they have to pay attention to detail across all of ARPI's categories. That they continuously ask "*what decision leads to the best possible result?*" It is the reality that experimental scientists face: there are a million ways to compromise an empirical study, and one has to avoid all of the pitfalls to achieve a meaningful answer.

ACKNOWLEDGMENTS

This work is part of a research program for teachers financed by the Netherlands Organisation for Scientific Research (Grant No. NOW—023.003.004).

[1] D. Hodson, Learning science, learning about science, doing science: Different goals demand different learning methods, Int. J. Sci. Educ. **36**, 2534 (2014).

[2] R. Millar, *The role of practical work in the teaching and learning of science, in Commissioned paper-Committee on High School Science Laboratories: Role and Vision* (National Academy of Sciences, Washington, DC, 2004).

[3] 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, Netherlands 1999), pp. 33–59.

[4] R. Gott and S. Duggan, A framework for practical work in science and scientific literacy through argumentation, Res. Sci. Technol. Educ. **25**, 271 (2007).

[5] R. Driver, P. Newton, and J. Osborne, Establishing the norms of scientific argumentation in classrooms, Sci. Educ. **84**, 287 (2000).

[6] 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, Netherlands 2012), pp. 189–207.

[7] S. E. Toulmin, *The Uses of Argument* (Cambridge University Press, Cambridge, England, 2003).

[8] S. Woolgar and B. Latour, *Laboratory Life: The Construction of Scientific Facts* (Princeton University Press, Princeton, NJ, 1986).

[9] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[10] B. J. Barron *et al.*, Doing with understanding: Lessons from research on problem-and project-based learning, J. Learn. Sci. **7**, 271 (1998).

[11] D. Hodson, Assessment of practical work, Sci. Educ. **1**, 115 (1992).

[12] P. Black and D. Wiliam, *Inside the Black Box: Raising Standards through Classroom Assessment* (Granada Learning, London, 2005).

[13] N. G. Holmes and C. Wieman, Introductory physics labs: WE CAN DO, Phys. Today **71**, 1 (2018).

[14] H. Lewandowski, *Colorado Learning about Science Survey for Experimental Physics (E-CLASS)* (APS, College Park, MD, 2014), p. S38.003.

[15] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).

[16] E. Etkina, A. Van Heuvelen, S. White-Brahmia, D. T. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 020103 (2006).

[17] 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, Philadelphia, 1991), pp. 167–178.

[18] R. Millar *et al.*, Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance, Res. Papers Educ. **9**, 207 (1994).

[19] 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**, 1 (2020).

[20] G. J. Kelly, Discourse practices in science learning and teaching, in *Handbook of Research on Science Education*, edited by N. G. Lederman and S. K. Abell (Routledge, 2014), Vol. II, Chap. 17, pp. 335–350.

[21] A. R. Cavagnetto, Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts, Rev. Educ. Res. **80**, 336 (2010).

[22] R. A. Duschl and J. Osborne, Supporting and promoting argumentation discourse in science education, Stud. Sci. Educ. **38**, 39 (2002).

[23] S. Erduran and M. P. Jiménez-Aleixandre, Argumentation in science education, in *Perspectives from Classroom-Based Research* (Springer, Dordrecht 2008).

[24] S. Erduran, J. Osborne, and S. Simon, The role of argumentation in developing scientific literacy, in *Research and the Quality of Science Education* (Springer, New York, 2005), pp. 381–394.

[25] S. Erduran, S. Simon, and J. Osborne, TAPping into argumentation: Developments in the application of Toul-

min's argument pattern for studying science discourse, Sci. Educ. **88**, 915 (2004).
[26] 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.
[27] 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.
[28] R. Gott and S. Duggan, Practical work: its role in the understanding of evidence in science, Int. J. Sci. Educ. **18**, 791 (1996).
[29] R. Roberts and P. Johnson, Understanding the quality of data: a concept map for 'the thinking behind the doing'in scientific practice, Curriculum J. **26**, 345 (2015).
[30] R. Roberts and C. Reading, The practical work challenge: incorporating the explicit teaching of evidence in subject content, Sch. Sci. Rev. **357**, 31 (2015), https://dro.dur.ac.uk/14738/2/14738.pdf?DDD29+ded4ss+d700tmt.
[31] J. Osborne, Teaching scientific practices: Meeting the challenge of change, J. Sci. Teach. Educ. **25**, 177 (2014).
[32] OECD, PISA 2015: DRAFT SCIENCE FRAMEWORK. 2013.
[33] R. Gott and R. Roberts, Concepts of evidence and their role in open-ended practical investigations and scientific literacy; background to published papers, in *The School of Education* (Durham University, UK, 2008).
[34] R. White and R. Gunstone, *Probing Understanding* (Routledge, London, 1992).
[35] S. Lachmayer, C. Nerdel, and H. Prechtl, Modelling of cognitive abilities regarding the handling of graphs in science education, Z. Naturwissenschaften **13**, 161 (2007).
[36] L. Cohen, L. Manion, and K. Morrison, *Research Methods in Education* (Routledge, London, 2013).
[37] S. Allen and J. Knight, A method for collaboratively developing and validating a rubric, Int. J. Scholarship Teach. Learn. **3**, n2 (2009).
[38] R. Kempa, *Assessment in Science* (Cambridge University Press, Cambridge, England, 1986).
[39] C.-C. Hsu and B. A. Sandford, The Delphi technique: Making sense of consensus, Pract. Assess. Res. Eval. **12**, 10 (2007).
[40] A. L. Delbecq, A. H. Van de Ven, and D. H. Gustafson, *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes* (Scott, Foresman, 1975).
[41] J. W. Murry Jr. and J. O. Hammons, Delphi: A versatile methodology for conducting qualitative research, Rev. High. Educ. **18**, 423 (1995).
[42] R. L. Custer, J. A. Scarcella, and B. R. Stewart, The Modified Delphi technique: A Rotational Modification, Int. J. Voc. Tech. Educ. **15** (1999).
[43] T. F. McNamara and T. J. Macnamara, *Measuring Second Language Performance* (Longman Publishing Group, London, 1996).
[44] D. M. Rubio, M. Berg-Weger, S. S. Tebb, E. S. Lee, and S. Rauch, Objectifying content validity: Conducting a content validity study in social work research, Social Work Research **27**, 94 (2003).
[45] L. Miller, Determining what could/should be: The Delphi technique and its application, in *Proceedings of the 2006 Annual Meeting of the Mid-Western Educational Research Association, Columbus, Ohio* (Midwestern Educational Research Association, Columbus, Ohio, 2006).
[46] M. Dancy, A. C. Lau, A. Rundquist, and C. Henderson, Faculty online learning communities: A model for sustained teaching transformation, Phys. Rev. Phys. Educ. Res. **15**, 020147 (2019).
[47] D. L. Gast, General factors in measurement and evaluation, in *Single Case Research Methodology* (Routledge, London, 2014), pp. 85–104.
[48] T. R. Kratochwill, *Single Subject Research: Strategies for Evaluating Change* (Academic Press, New York, 2013).
[49] G. Welford, W. Harlen, and B. Schofield, *Assessment of performance unit: Practical testing at ages 11, 13 and 15* (Department of Education and Science, London, 1985).
[50] S. Johnson, G. Britain, and S. A. o. P. Unit, *National Assessment: the APU science approach*, (HM Stationery Office, London, 1989).
[51] R. Millar and J. Osborne, *Beyond 2000: Science Education for the Future* (King's College London, School of Education, 1998), https://www.nuffieldfoundation.org/wp-content/uploads/2015/11/Beyond-2000.pdf.
[52] National Research Council, *Next Generation Science Standards: For States, by States* (National Academies Press, Washington, DC, 2013).
[53] R. Millar, Taking scientific literacy seriously as a curriculum aim, in *Asia-Pacific Forum on Science Learning and Teaching* (The Education University of Hong Kong, Hong Kong, 2008).
[54] European Commission, White paper on education and training. Teaching and Learning: Towards the learning society, Brussels (1995), retrieved from https://op.europa.eu/nl/publication-detail/-/publication/d0a8aa7a-5311-4eee-904c-98fa541108d8/language-en.
[55] D. A. Roberts and R. W. Bybee, Scientific literacy, science literacy, and science education, in *Handbook of Research on Science Education*, edited by N. Lederman and S. K. Abell (Routledge, New York, 2014).
[56] W. Ottevanger *et al.*, Kennisbasis natuurwetenschappen en technologie voor de onderbouw vo: Een richtinggevend leerplankader, SLO (nationaal expertisecentrum leerplanontwikkeling) (2014).
[57] H. Eijkelhof, Curriculum policy implications of the PISA scientific literacy framework, in *Electronic Proceedings of the ESERA 2013 Conference, Strand 10, Science Curriculum and Educational Policy* (ESERA, Nicosia, Cyprus, 2014).
[58] Netherlands Institute for Curriculum Development, Retrieved from http://international.slo.nl (2016).
[59] S. Breakspear, The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance (2012).
[60] S. G. Sunder, *Connecting IB to the NGSS: The Dual Implementation of International Baccalaureate and the Next Generation Science Standards: Challenges and*

Opportunities, edited by I. B. Organization (International Baccalaureate Organization, Geneva, 2016).
[61] M. o. E. Singapore, *Physics Syllabus Pre-University*, edited by C. P. a. D. Division (2019).
[62] Ministry of Education Singapore, *Science Syllabus Lower, and Upper Secondary*, edited by M. o. Education (2013), p. 76.
[63] N. Burdett and L. Sturman, A Comparison of PISA and TIMSS against England's National Curriculum, in *Proceedings of the 5th IEA International Research Conference* (2013).
[64] E. F. Redish and J. S. Rigden, *The Changing Role of Physics Departments in Modern Universities: Proceedings of ICUPE* (AIP Press, New York, 1998) [Imprint].
[65] J. Kozminski *et al.*, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (American Association of Physics Teachers College Park, MD, 2014), p. 29.
[66] A. Van Heuvelen, Learning to think like a physicist: A review of research-based instructional strategies, Am. J. Phys. **59**, 891 (1991).
[67] C. Wieman, Comparative cognitive task analyses of experimental science and instructional laboratory courses, Phys. Teach. **53**, 349 (2015).
[68] E. Rusman and K. Dirkx, Developing rubrics to assess complex (generic) skills in the classroom: How to distinguish skills' mastery levels?, Pract. Assess. Res. Eval. **22**, 12 (2017).
[69] B. M. Moskal, Scoring rubrics: What, when and how?, Pract. Assess. Res. Eval. **7**, 3 (2000).
[70] S. M. Brookhart, *The Art and Science of Classroom Assessment. The Missing Part of Pedagogy. ASHE-ERIC Higher Education Report* (ERIC, Washington DC, 1999), Vol. 27, No. 1.
[71] J. S. Lederman, N. G. Lederman, S. A. Bartos, S. L. Bartels, A. Antink Meyer, and R. S. Schwartz, Meaningful assessment of learners' understandings about scientific inquiry—The views about scientific inquiry (VASI) questionnaire, J. Res. Sci. Teach. **51**, 65 (2014).
[72] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.18.010111 for the assessment rubric for physics inquiry.
[73] J. Biggs, Enhancing teaching through constructive alignment, Higher Educ. **32**, 347 (1996).
[74] 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, http://www.iupap-icpe.org/publications/teach1/ConnectingResInPhysEducWithTeacherEduc_Vol_1.pdf.
[75] D. R. Dounas-Frazer and H. Lewandowski, The modelling framework for experimental physics: Description, development, and applications, Eur. J. Phys. **39**, 064005 (2018).
[76] B. M. Zwickl, D. Hu, N. Finkelstein, and H. J. Lewandowski, Model-based reasoning in the physics laboratory: Framework and initial results, Phys. Rev. ST Phys. Educ. Res. **11**, 020113 (2015).