Development of a teaching-learning sequence for scientific inquiry through argumentation in secondary physics education

Pols, C.F.J.

DOI
10.4233/uuid:df26ce2c-4b0f-41ee-93ff-301aa82457c3

Publication date
2023

Document Version
Final published version

Citation (APA)
Pols, C. F. J. (2023). *Development of a teaching-learning sequence for scientific inquiry through argumentation in secondary physics education*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:df26ce2c-4b0f-41ee-93ff-301aa82457c3

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Development of a teaching-learning sequence for scientific inquiry through argumentation in secondary physics education

Freek Pols

Development of a teaching-learning sequence for scientific inquiry through argumentation in secondary physics education

Development of a teaching-learning sequence for scientific inquiry through argumentation in secondary physics education

Proefschrift
ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 17 januari 2023 om 10:00 uur

door
Christianus Fredericus Johannus POLS
Master of Science in Applied Physics, Delft University of Technology
Geboren te Lelystad, Nederland

Dit proefschrift is goedgekeurd door de promotor: prof. dr. M.J. de Vries

Samenstelling promotiecommissie:

Rector Magnificus	Voorzitter
Prof. dr. M.J. de Vries	Technische Universiteit Delft, promotor
Dr. P.J.J.M. Dekkers	Technische Universiteit Delft, copromotor

Onafhankelijke leden:

Prof. dr. M. de Cock	Katholieke Universiteit Leuven, België
Prof. dr. A. M. Pendrill	Universiteit van Gotenburg, Zweden
Prof. dr. ir. C.R. Kleijn	Technische Universiteit Delft
Prof. dr. J.T. van der Veen	Techniche Universiteit Eindhoven

Reserve lid:

Prof. dr. J.M. Thijssen	Technische Universiteit Delft

This research was funded by the Netherlands Organization for Scientific Research (NWO) (Project no. 023.003.004)

Keywords: physics education, scientific inquiry, argumentation, concepts of evidence, practical work

Cover drawing made by Marjolein ter Braak
Printed by: ProefschriftMaken.nl || www.proefschriftmaken.nl

Table of contents

1. Introduction

Almost everyone, at some point in secondary school physics education, has conducted an experiment. Conducting experiments is seen as a valuable, if not essential, part of physics education. However, such activities are also costly and time-consuming. Moreover, it is often reported that these activities hardly contribute to the students' content knowledge, nor that they enhance students' view on how science works or teaches them how to plan an inquiry independently (Abrahams, 2011; Abrahams & Millar, 2008; Hodson, 1990, 2014; Hofstein & Kind, 2012; Hofstein & Lunetta, 1982, 2004; Lunetta, Hofstein, & Clough, 2007). The aforementioned problems have been the point of departure for developing a teaching-learning sequence (TLS) that develops students' understandings of doing physics inquiry and studying how we could effectively enable students to engage in physics inquiry. Although developing a deeper understanding of physics inquiry is an important goal of physics education on its own, it is my contention that a deeper understanding of physics inquiry also contributes to the improvement of the learning outcomes of experiments that focus on developing physics content knowledge in students. I discuss this in more detail below.

Above I used the casual term conducting experiments. However, in literature one can find many related terms such as *doing science* / *practical science* / *scientific inquiry* / *experimental inquiry* or *enquiry* / *research* / *labs* or *lab work* and so on (Millar, 2015). In European literature, the term practical work – as opposed to lab work – is frequently used to denote that the activities in which students manipulate and observe real objects and materials can be conducted outside the laboratory (Millar, Le Maréchal, & Tiberghien, 1999). I will use the term practical work as the umbrella term for these activities, the term experiment to denote the actual experiment that is carried out (often in the classroom in order to collect data) and the term inquiry as the whole process of conceiving, conducting and evaluating an experiment where the main learning goal relates to learning how to do scientific research.

1.1 Problem definition

Practical work is part of physics education because physics is an empirical discipline: our knowledge of nature is developed through experimental work and otherwise only of true value when confirmed experimentally. *If theory does not agree with experiment, theory is simply wrong* (Feynman Cornell University Lecture, 1964, reported in Feynman, Leighton, & Sands, 2011). That is, of course, on the premise that the experiment is adequately carried out. That reason alone – the empirical nature of physics – suffices to justify conducting practical work in secondary school science education (Millar et al., 1999).

A second justification is the idea that the development of content knowledge is enhanced when students have hands-on experience with the content (Millar, 1991; Millar

et al., 1999; Woolnough & Allsop, 1985). Abstract physics becomes more tangible when students have a feel for the phenomenon, when a direct link between theory and practice is made, see Figure 1.1.

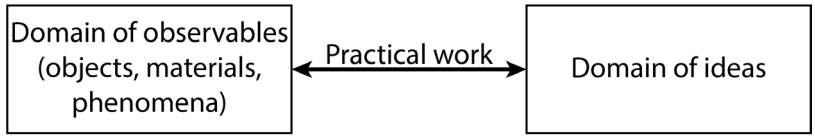


Figure 1.1: Practical work serves as a link between theory (ideas and concepts) and practice (observation and data) (Millar, 1999).

A third justification is that it contributes to raising *scientific literacy* (Gott & Duggan, 2007; Kanari & Millar, 2004; Millar et al., 1999), the general scientific awareness and understanding of science required to participate meaningfully in modern societal issues. According to PISA 2015 (OECD, 2013), to become a scientifically literate person one has to develop three core abilities:

(1) Explain phenomena scientifically: Recognise, offer and evaluate explanations for a range of natural and technological phenomena.
(2) Evaluate and design scientific inquiry: Describe and appraise scientific investigations and propose ways of addressing questions scientifically.
(3) Interpret data and evidence scientifically: Analyse and evaluate data, claims and arguments in a variety of representations and draw appropriate scientific conclusions.

It is not difficult to see how core abilities (2) and (3) can be developed using practical work (Gott & Duggan, 2007; Hofstein & Kind, 2012; Kanari & Millar, 2004; Millar et al., 1999).

A fourth justification is that doing practical work prepares students for more complex and independent inquiries in tertiary education. In doing practical work, students develop expertise in doing scientific inquiry. Many of the required understandings to successfully engage in scientific inquiry can be taught, developed and applied in secondary school science practical work in which students are introduced to the methods of science (Hodson, 2014; Millar, 1991; Tamir, 1991).

However, these justifications may need further scrutiny as in the extensive literature on this topic it is often reported that the associated learning goals are hardly attained using practical work (Abrahams, 2011; Abrahams & Millar, 2008; Hodson, 1990, 2014; Hofstein & Kind, 2012; Hofstein & Lunetta, 1982, 2004; Lunetta et al., 2007). Despite many decades of research and development, educational reforms and various shifts in educational focus of practical work, the learning outcomes of practical work remain disappointing, both in terms of developing students' ability to conduct scientific inquiry independently and their understanding of the relations between scientific theory and the empirical world. Research has only been able to demonstrate that practical work is superior to other teaching methods

in teaching how to manipulate equipment. Students frequently do not know what the purpose of the experiment is, carry out experiments that are at the best pseudo-scientific, only draw superficial or even faulty conclusions and/or do not learn what the teacher intended the students to learn (Abrahams, 2005; Abrahams & Millar, 2008; Hofstein & Lunetta, 2004; Jenkins, 1998; Lunetta et al., 2007; Tamir, 1991; van den Berg, 2013; Watson, Goldsworthy, & Wood-Robinson, 1999). Viewed from this perspective – acknowledging the limited learning outcomes of practical work, the heavy burden on students' time, the teacher's energy put into these activities, the specially equipped classrooms and the school's limited budget – practical work can hardly be justified (Abrahams & Millar, 2008; Hodson, 1990, 1991, 2014; Hofstein & Lunetta, 1982; Jenkins, 1998; Tamir, 1991). Hodson (1991) writes:

> *As practised in many schools, practical work is ill-conceived, confused and unproductive. For many children, what goes on in the lab contributes little to their learning of science or to their learning about science and its methods. Nor does it engage them in doing science in any meaningful sense. At the root of the problem is the unthinking use of lab work.*

Hodson refers here to teachers' unthinking use of practical work in their science lessons for all kinds of purposes. But in addition to the way teachers use practical work, there are issues with the way learners conduct it as well. Millar et al. (1999, p. 34), one of the leading experts in the field of practical work in science education, adds:

> *Frequently practical work is carried out very rapidly, or with unreliable equipment, or with insufficient attention to care and precision, so that students fail even to produce the phenomenon they are supposed to observe, let alone be helped to appreciate patterns, trends or explanations. Even when the outcomes are as the teacher intended, conclusions which seem 'obvious' to the teacher can appear less so to the student.*

These statements are more than two decades old, but more recent studies have shown that not much has changed since then (Abrahams & Millar, 2008; Abrahams & Reiss, 2012; Hodson, 2014; Hofstein, 2017; Hofstein & Kind, 2012; Hofstein & Lunetta, 2004; Holmes, Olsen, Thomas, & Wieman, 2017; Holmes & Wieman, 2018; Kok, Priemer, Musold, & Masnick, 2019; Lunetta et al., 2007; Wieman, 2015). As summarised by Holmes (2018):

> *The only thinking the students said they did in structured and content-focused labs (the kind in our study of nine courses) was in analysing data and checking whether it was feasible to finish the lab in time.*

My experience as a physics teacher is that students happily conduct the experiments they are given, follow the stepwise instructions and in following these descriptions succeed

to some degree in confirming the scientific law that was under investigation. However, only a few students really understand how the experiment relates to the theory covered in the previous lesson and appreciate the connection between the theoretical content and the practical work. Most students are actively engaged in getting the job done, but do not persevere to achieve a scientifically acceptable outcome. Students are working hands-on rather than minds-on (Abrahams & Millar, 2008; van den Berg, 2013).

Despite these objections to practical work in secondary school science education, most teachers would not abandon practical work as an educational tool. Many science education researchers and teacher educators still believe in the potential value of practical work. As a teacher, I too believe in the added value of practical work. I believe that practical work can and must be improved. As a researcher, I have the ability and responsibility to seek ways of improving and reflecting on new ways of conducting practical work. As teacher-researcher I can unite the best of both worlds by designing, implementing, testing, reflecting on and improving practical work. This study focusses on increasing the learning outcomes of practical work in secondary school physics education by developing students' knowledge of doing physics inquiry and exploring how we can do so effectively.

I started this doctoral study with the assumption that when students have more knowledge about collecting, processing and analysing data this will result in an improvement of the general quality of students' practical work: Developing conceptual knowledge through practical work, e.g., having students independently conclude that there is a proportional relationship between voltage and current when using an Ohmic resistor, requires firstly the collection of accurate and reliable data. This necessitates that students, e.g., carefully choose the equipment, and use a range and interval that reveals the pattern and its details. Secondly, it requires an ability to process and analyse the data, and an understanding of measurement uncertainty. Hence, if practical work is used to enhance conceptual learning, students should have a basic knowledge about the validity and reliability of data.

In my teaching, I have designed and used various activities to enhance this knowledge so students at least know that, e.g., measurements should be repeated (Pols, 2016, 2017, 2020b). However, as it turned out, knowledge about collecting accurate and reliable data alone was not sufficient to have students successfully carry out practical work and produce a scientifically acceptable answer to the research question (or often even an answer that is acceptable for the teacher). The mistake I, as a teacher, probably made is that I saw students as already part of the critical community of scientists having, in the words of Oreskes (2019) – who uses Popper's ideas – *an attitude of scepticism and disbelief*, often referred to as a critical attitude or scientific attitude. I did tell them *what* to do but probably failed to adequately explain *why*… As a consequence, if not urged by the teacher, students will seldom think about what they are doing and why in that particular way. They do not consider better ways to collect, process or analyse data. Rather than blaming our

students for their lack of critical attitude, this notion is meant as a clear statement of how things are. There is no incentive for students to use their acquired inquiry knowledge – or more precisely, to use their *conceptual and procedural knowledge* (see theoretical background section 1.2.2) (Millar, 1997). Acknowledging this might help in developing activities that foster students' critical attitude during practical work.

In the intervention studies carried out, I sought ways to produce an incentive strong enough to make students think about the best way to collect data and evaluate the quality of their work. The idea of '*striving for scientific cogency',* convincing yourself and subsequently others that what you have found provides the best possible answer to the given research question because it is as reliable, useful and informative as possible, came up as a central theme to develop students' understanding to carry out physics inquiries independently. In a scientific approach, researchers evaluate critically with every step they make whether the chosen option (e.g. instrument use; range etc.) is the best available within given constraints. If students understand and experience the necessity of making use of this approach, it is more likely they are motivated to collect accurate data. This requires them to understand what makes data reliable and when data can be considered evidence. This, in turn, might result in students paying attention during activities in which this knowledge is developed. If students feel that they ought to convince others that the best available answer has been produced, they would feel required to engage in argumentation from evidence, including data analysis and interpretation.

1.2 Theoretical background

The ideas of *argumentation*, the *PACKS model* as a suitable model for thinking about and reflecting on the teaching and learning taking place during practical work, and the *Concepts of Evidence* (CoE) form the underpinning theoretical notions of this research. It is thus important to briefly elaborate on these ideas individually, and then show how these are combined to form the backdrop of this research.

1.2.1 Argumentation in scientific inquiry

To better understand how we can teach students to plan and carry out their own physics inquiry and evaluate the quality of the work of others, it might be useful first to look at how experimental physicists approach their inquiries. After all, they are the experts, and a major aim of physics education is to teach students to think like a physicist (Etkina, 2015).

An experimental physicist might regard a scientific inquiry as the building of a scientifically cogent argument for a specific claim. At the outset of the inquiry, the researcher does not know precisely what claim (conclusion) it will yield. However, regardless of the precise outcome of the inquiry, the researcher tries to make the future claim as indisputable as possible, defending it against any potential criticism. The researcher

collects data in a reliable and valid manner, uses underlying theories to support the method and ideas, interprets data, weighs evidence, assesses alternative methods and explanations of the observed, makes claims and sets limits to the conclusions and its validity. These are all components of a scientific argument (Gott & Duggan, 2007; Toulmin, 2003; Woolgar & Latour, 1986).

Once the researcher has produced an answer to the research question, argumentation is used to elaborate the (new) scientific ideas, which are further elucidated in scientific articles and at conferences. Argumentation is used by peers to dispute or accept these ideas and claims. As Driver, Newton, and Osborne (2000, p. 288) state: *It is through such processes of having claims checked and criticized that 'quality control' in science is maintained.*

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* (Kelly, 2014, p. 329), is thus an activity that lies at the heart of science and scientific inquiry. Teaching students how to plan, carry out and evaluate a rigorous inquiry involves teaching them argumentation:

> *In doing science, students have responsibility for posing questions, devising methods of inquiry, analysing and interpreting data, reaching a conclusion, constructing a convincing argument for that conclusion, and communicating their methods, findings and conclusions to others (Hodson, 2014).*

1.2.2 The structure and content of an argument

A useful tool think about argumentation, both its structure (field-invariant elements) and its content (field-dependent elements), is the Toulmin (2003) argumentation model. For secondary science inquiry, Gott & Duggan (2007) adapted the Toulmin argumentation model to (secondary) science inquiry, see figure 1.2. The experimental physicist (investigator) draws a *claim* that is based on and supported by *data* and connected through *warrants*: the reasoning that defends the claim based on the data. These warrants can be further supported by *backings*. These backings are, according to the authors, *the detailed facts and ideas which underpin the data collection such as the number of readings taken, the method of averaging, the validity of the measurement itself and so on*. The claim is further strengthened when *qualifiers* and *rebuttals* are included. These field-independent elements set limitations to the validity of the claim and defend it against any potential criticism.

According to Toulmin an argument consists at least of the field-invariant elements data, warrants and claim. The argument is further strengthened by including qualifiers and rebuttals. However, the true strength, or scientific cogency, of the argument depends on the argument's content. Toulmin (2003, p.137) states: *the standards for judging the soundness, validity, cogency or strength of arguments are in practice field-dependent*.

Whether data is accepted as evidence (what the underlying rules are), and whether the claim is substantially and sufficiently supported by the data, is field-dependent. In order to enable students to engage independently in scientific inquiry we should at least teach students what these field-dependent elements are. But so far, attention for the field-dependent elements of argumentation in teaching scientific inquiry remains underexposed.

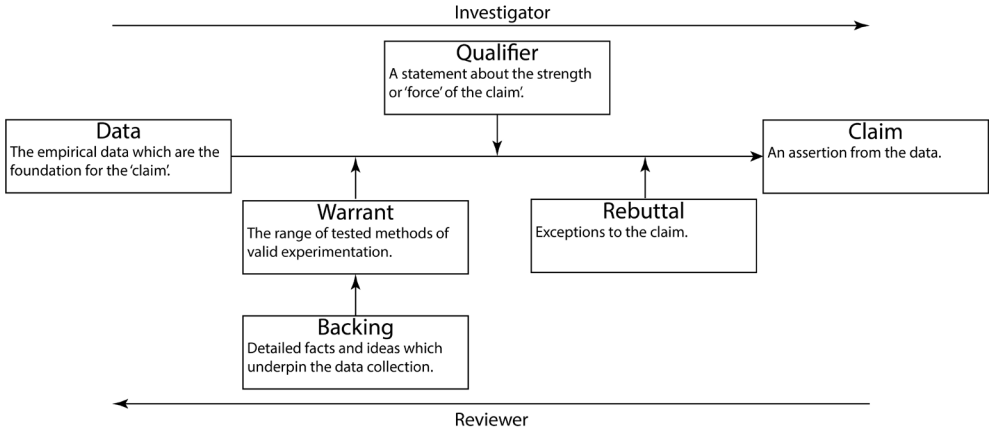


Figure 1.2 The Toulmin (2003) argumentation model as adapted by Gott & Duggan (2007) to (secondary) science inquiry.

1.2.3 Argumentation in teaching scientific inquiry

Despite the fact that argumentation is an integral part of scientific inquiry, and various science education researchers have emphasized the importance of argumentation in science education, argumentation is only scarcely utilized in science classrooms (Cavagnetto, 2010; Driver et al., 2000; Duschl & Osborne, 2002; Erduran & Jiménez-Aleixandre, 2008; Erduran, Osborne, & Simon, 2005; Erduran, Simon, & Osborne, 2004; Hofstein & Kind, 2012; Newton, Driver, & Osborne, 1999; Osborne & Dillon, 2008). Students are hardly ever asked to show that their answer is the best answer obtainable (Gunstone & Champagne, 1990; Hodson, 1990, 2014; Holmes & Wieman, 2016, 2018; Tamir, 1991; Wieman, 2015). Indeed, when observing students who engage in practical work one might have the impression that students are not concerned with producing a scientifically convincing answer that is supported by the data and which is thoroughly substantiated with arguments. It seems as if practical work is mostly a hands-on activity rather than an act of mind. However, successful engagement in practical work requires that students know what they are doing, and why. In other words, it requires a substantial amount of procedural and conceptual knowledge (Gott & Duggan, 1995; Millar, 1997; Millar, Lubben, Gott, & Duggan, 1994). To understand what this procedural and conceptual knowledge is, we elaborate on the *PACKS model*.

1.2.4 The PACKS Model

Millar et al. (1994) regard practical work as a knowledge-based activity and demonstrate that the application of more relevant knowledge leads to better performance. The influence of knowledge on students' performance is illustrated in Millar et al.'s *Procedural and Conceptual Knowledge in Science* (PACKS)-model (Figure 1.3) in which a link is made between the various types of knowledge and their influence in each step of a scientific inquiry. Students' knowledge influences, amongst others, how the task is understood, how the inquiry is set up, which instruments are chosen, how these are used, how often measurements are repeated and what measurement range and interval are chosen. Each decision influences the next steps made in the inquiry:

> *Thus, almost every move that a scientist makes during an inquiry changes the situation in some way, so that the next decisions and moves are made in an altered context. Consequently, scientific inquiry is holistic, fluid and reflexive, not a matter of following a set of rules that requires particular behaviour at particular stages. It is an organic, dynamic, interactive activity, a constant interplay of thought and action (Hodson, 2014).*

This implies that the PACKS model is a non-linear model: there are several back-loops. For instance, the evaluation of the accuracy of the data obtained might lead to a revised method and the collection of more or alternative data. In the PACKS model vectors indicating these back-loops have been omitted.

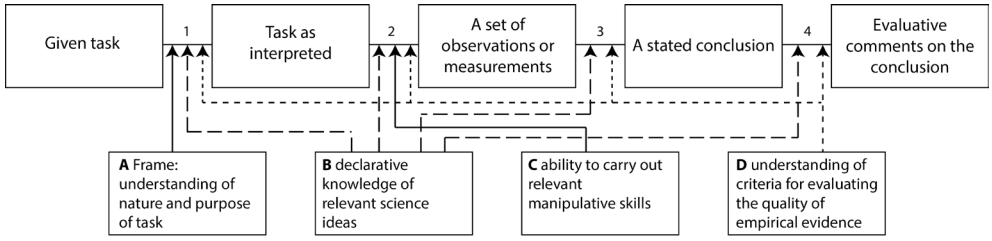


Figure 1.3: The PACKS-model relates different types of knowledge pertaining to the various decisions made during different phases in a scientific inquiry (Millar et al., 1994)

In the PACKS-model, knowledge domains **A**, **C** and **D** are categorized as *procedural knowledge*: *the understandings which inform actions in response to practical investigation tasks in science in general* (Millar et al., 1994). Especially knowledge domain **D**, *understanding of criteria for evaluating the quality of empirical evidence*, crucially influences the quality of students' work. Gott and Duggan (1995) operationalized the criteria for evaluating the quality of empirical evidence by formulating *Concepts of Evidence*: concepts underpinning the ideas related to the reliability and validity of scientific inquiry and scientific evidence.

1.2.5 Concepts of Evidence

A concept of evidence (CoE) is defined by Gott and Duggan (1995, p. 30) *as an idea that draws attention to the importance of procedural understanding and the concepts underlying the doing of science in relation to the evidence as a whole.* Their first set of twelve CoE has been extended to a comprehensive but tentative list of almost one hundred concepts such as fair testing, repeatability, reproduction, measurement uncertainty, accuracy and precision, see Table 1.1 (Gott, Duggan, Roberts, & Hussain, 2003). Gott and Duggan (2007) argue that these CoE are the common basis of various types of inquiry as the CoE relate to the design of the inquiry, collection and processing of data and the evaluation of the inquiry as a whole in terms of reliability and validity. In almost all inquiries one needs to understand the ideas that underpin scientific evidence: what is a single datum or observation, what is meant by a reliable dataset, and how can we determine relations between variables and draw a conclusion from these? The CoE play a central role in assessing the reliability and validity of scientific inquiries of others as well. A reviewer uses concepts such as *fair test* – *in which variables are controlled to isolate the effect of the independent variable on the dependent variable* – to judge the evidence and provide an argument whether to accept or decline the claim.

If we want to enable students to independently carry out scientific inquiries, we should develop students' understandings of the CoE (Gott & Duggan, 1995, 2003, 2007; Gott et al., 2003; Millar, 1997; OECD, 2013; Osborne, 2014b; R. Roberts, Gott, & Glaesser, 2010). Knowledge of the CoE is required to decide, e.g., whether a measurement should be repeated or whether the right choice for an instrument is made given the required accuracy of the data. Referring to these concepts, Millar et al. (1994) already argued almost three decades ago that:

> *There is a need to devise activities which progressively develop and refine children's understanding of the purpose of scientific investigation, and of the key concepts which underpin judgements about the quality of data. Successful performance of a scientific investigation requires the ability to plan the collection of data, to collect and interpret real data, and to make and defend judgements on the basis of the whole data set collected, informed by an appropriate assessment of the reliability of the data.*

As the CoE are well defined, these concepts can be taught and students' understanding of these assessed. Once the (basic) CoE are understood by the students, these can be used as 'tools' from an extensive toolkit. If a basic understanding is present, these 'tools' help make the right choices in every step of a scientific inquiry (Millar, 1997). Students extend their toolkit in new, more complex inquiries (T. Smits, Lijnse, & Bergen, 2000). They learn how to *do science* in the best, most authentic way: by *doing science* (Hodson, 2014).

Table 1.1: A selection of CoE with the concepts on the left and a brief explanation of the concept on the right (Gott et al. 2003)

Concept	Understanding that ….
10. Resolution and error	... the resolution is the smallest division which can be read easily. The resolution can be expressed as a percentage.
18. Trueness or accuracy	... trueness is a measure of the extent to which repeated readings of the same quantity give a mean that is the same as the 'true' mean.
21. Reproducibility	... whereas repeatability (precision) relates to the ability of the method to give the same result for repeated tests of the same sample on the same equipment (in the same laboratory), reproducibility relates to the ability of the method to give the same result for repeated tests of the same sample on equipment in different laboratories.
27. Range	... the range is a simple description of the distribution and defines the maximum and minimum values measured.
46. Fair test	... a fair test is one in which only the independent variable has been allowed to affect the dependent variable.
60. Reliability of the design	... the reliability of the design includes a consideration of all the ideas associated with the measurement of each and every datum.
87. Practicality of consequences	… the implications of the evidence may be practical and cost effective, or they may not be. The more impractical or costly the implications, the greater the demand for higher standards of validity and reliability of the evidence.

Although Gott and Duggan (2007) argue that the CoE may form the basis for setting up scientific inquiries, in the end, students have to provide a convincing answer to the research question. To ensure that a scientifically cogent answer is produced, students have to think from the start of their inquiry about how to substantiate the answer that is still unknown to them. This requires a combination of knowledge of CoE, the use of the various PACKS types of knowledge, a thorough understanding of the purpose of scientific inquiry, and the use of argumentation.

1.2.6 Combining scientific inquiry, argumentation and concepts of evidence

The possibility and necessity of combining argumentation and scientific inquiry has been reported by many scholars (Driver et al., 2000; Gott & Duggan, 2007; Hofstein & Kind, 2012; Kim & Song, 2006; Sampson, Grooms, & Walker, 2009, 2011; Walker, Sampson, & Zimmerman, 2011). Even the combination of scientific inquiry, argumentation and concepts of evidence have been introduced, and activities to foster argumentation have been proposed (Gott & Duggan, 2007). However, at the outset of this study it was unclear – from a theoretical perspective – whether and how the Toulmin argumentation model and the

PACKS model can be combined or how these models complement each other. The field-invariant elements in the Toulmin argumentation model are well-known, but what are these field-dependent elements, in which the CoE surely play an important role? In the theoretical section above, we illustrate that there seems to be a connection between the Toulmin argumentation model, the PACKS model and the Concepts of Evidence, but a clear connection between these elements is hitherto lacking. Moreover, the implementation of these ideas and operationalization of such proposed activities have not been substantially studied empirically. Ergo, from a practical perspective there remain questions of how to enable students to engage in scientific inquiry and integrate argumentation in scientific inquiry in secondary school science education. What do students need to know and understand about doing scientific inquiry? What is the role of the concepts of evidence in this? How do we teach these understandings effectively? How can we optimize such activities so that students learn how to plan a rigorous inquiry? What design principles help in constructing such activities, and why do these help (or in what way)? And last but not least: How do we motivate students to learn all this?

1.3 Research scope and questions

The scope of this study is limited to physics experiments in which students establish a mathematical relationship between two variables. We here introduce the term *quantitative physics inquiries* (QPI) to refer to this type of experiment. Although QPI does not include all types of experiments, it includes the vast majority of physics experiments carried out at secondary school (Henderson, 1996; McDermott & Redish, 1999). We investigate how we can engage students in planning, conducting and evaluating rigorous QPI. In this context we mean by *rigorous* that the inquiry might be simple content wise, but from a scientific perspective students take well-considered and justifiable choices regarding the design, method and procedures in their inquiry.

If we want to enable students to independently plan, carry out, and report a rigorous QPI, the first question to be addressed is what this precisely entails, and whether students already acquired part of the knowledge involved. Therefore the first two central questions addressed in this thesis are:

> *1 What knowledge about scientific inquiry is required to plan, carry out and report a rigorous quantitative physics inquiry and how can mastery of this knowledge be assessed?*
>
> *2 What part of this knowledge has been thoroughly acquired by students who enter upper secondary school?*

From here on we will call the knowledge referred to in these research questions: inquiry knowledge. If we know more precisely what inquiry knowledge students have, or lack, we can build activities that constitute a TLS that targets these knowledge deficiencies. Developing such activities is more effective if we can rely on design principles that encourage students to engage in rigorous QPI and result in activities that foster attainment of the learning goals. Therefore the subsequent two central questions addressed in this thesis are:

> *3 What design principles are effective in guiding the design of a teaching-learning sequence that aims at enhancing students' critical attitude and developing inquiry knowledge?*
>
> *4 What do students learn in a teaching-learning sequence directed at teaching inquiry through argumentation in terms of inquiry knowledge, enhanced critical attitude and use of argumentation?*

Combining the theoretical framework, the four research questions above and our broad aim to engage students in more authentic QPI independently, the overarching research question of this thesis is:

> *What, and how, does paying attention to argumentation in inquiry contribute to enabling students to successfully engage in quantitative physics inquiry?*

1.4 Outline

This thesis consists of seven research related chapters. The remaining six chapters consist of four research chapters, a chapter elaborating the developed TLS and one concluding chapter. The links between the research chapters and the research questions are summarized in Table 1.4. In chapter 2 we describe what students are expected to know about analysing empirical data when entering upper secondary education (in the Netherlands), and whether the participating students acquired a proper attainment level. Chapter 3 reviews what devising a rigorous QPI entails and how we can assess students' attainment level regarding the associated inquiry knowledge. Chapter 4 describes the developed TLS in more detail. We do not regard this as a research chapter, but it provides more details of the TLS. In chapter 5 we elaborate on the first activity of this TLS. The activity is used to make students aware of the importance of upholding scientific standards and creating circumstances where students *want* to find a scientifically acceptable answer to the research question. Furthermore, the study provides insights in what students know and do when asked to carry out a physics inquiry. In chapter 6 we elaborate on the effectiveness of the TLS with regard to the development of students' understanding and the development of their critical attitude towards physics inquiry. We review how our design principles

contribute to making the TLS effective. In chapter 7 we answer the four research questions and the main question. We elaborate, briefly, on the implications of study this study with regard to physics education and provide directions for future research.

Due to the article-based nature of this dissertation and the wish to have the studies in each of these chapters to be read independently, the introductions and theoretical framework in the various chapters partly overlap. Moreover, throughout the studies we make extensive use of abbreviations. A list of abbreviations and their meaning is provided in section 12.1 on page 184.

Table 1.2: An overview of the links between the chapters and the three research questions.

Ch.	Study	Shirt title	RQ 1	RQ 2	RQ 3	RQ 4
2	1	Investigating students' ability to analyse experimental data in secondary physics education	*			
3	2	Defining and assessing understandings of evidence with ARPI	*			
4	3&4	The development of the teaching-learning sequence			*	
5	3	Fostering a scientific approach to secondary physics inquiry	*	*	*	
6	4	Developing students' understanding of evidence		*	*	*
7		General conclusions	*	*	*	*

2. What do they know? Investigating students' ability to analyse experimental data in secondary physics education

Article previously published as:

Pols, C.F.J., Dekkers, P.J.J.M., and de Vries, M.J. (2021). What do they know? Investigating students' ability to analyse experimental data in secondary physics education. *International Journal of Science Education,* DOI: 10.1080/09500693.2020.1865588

This paper explores students' ability to analyse and interpret empirical data as inadequate data analysis skills and understandings may contribute to the renowned disappointing outcomes of practical work in secondary school physics. Selected competences, derived from a collection of leading curricula, are explored through interviews and practical tasks, each consisting of three probes. The 51 students, aged 15 and commencing post-compulsory science education in the Netherlands, were able to carry out basic skills such as collecting data and representing these. In interpreting the data in terms of the investigated phenomenon or situation however, performance was weak. Students often appeared to be unable to identify the crucial features of a given graph. Conclusions based on the data were often tautological or superficial, lacking salient features. Students failed to infer implications from the data, to interpret data at a higher level of abstraction, or to specify limitations to the validity of the analysis or conclusions. The findings imply that the students' understanding of data-analysis should be developed further before they can engage successfully in more 'open' practical work. The study offers a collection of activities that may help to address the situation, suggesting a baseline for guided development of data analysis abilities.

2.1 Introduction

We expect students who have completed the compulsory part of science education to be able to carry out a basic quantitative physics inquiry (QPI), in which a relationship between two variables is determined, independently. Focusing on data analysis and interpretation, we explore whether this expectation is justified for Dutch students, and how potential deficiencies may be diagnosed and eventually addressed.

The common approach to educational aims related to scientific inquiry is through *practical work* (Millar, 2010), activities in which students observe or manipulate the objects of interest (Millar et al., 1999), and draw conclusions from the data collected (Kanari & Millar, 2004). The main purpose of practical work is usually to 'discover' physics concepts or enhance conceptual knowledge by establishing the relationship between physical quantities (Hofstein, 2017; Millar et al., 1999). Although an important part of finding and interpreting these relationships involves carrying out a proper data analysis, enhancing competence in data analysis is rarely the central objective of practical work. Still, students often encounter various insuperable problems when analysing data resulting in superficial and incomplete conclusions (Kanari & Millar, 2004) or even 'alternative science' (Hodson, 1990). A lack of competence in data analysis potentially contributes to the limited learning outcomes of practical work that prompt some to wonder if the same learning goals may and should be achieved with less costly and time consuming methods (Hodson, 1990; Hofstein, 2017; Hofstein & Kind, 2012; Hofstein & Lunetta, 2004; Lunetta et al., 2007; van den Berg, 2013).

Guided by the central research questions:

1 How do 15-year old students, after completing compulsory science education and entering a pre-academic science-based exam program analyse experimental data?

2 What is the quality of that analysis?

this study investigates whether students who have just finished the compulsory part of science education in the Netherlands have the ability to analyse and interpret experimental data by constructing adequate data representations and drawing qualified, appropriate, defensible conclusions from these data. From this baseline, potential deficiencies may be specified as a starting point for designing suitable learning pathways to develop more advanced kinds of QPI later.

2.2 Background

2.2.1 Dutch educational system

In the Netherlands, students enter one of three ability levels of lower secondary education at age 12. The stream for preparatory vocational education (opted for by app. 60% of the Dutch student population) is not considered here, as the study takes place in a school preparing for higher vocational education (opted for by app. 20%) or university studies (opted for by app. 20%) (DUO, 2017). After three years of lower secondary education with physics as a compulsory subject in the second and third year, students choose between a program oriented towards the natural *sciences* and towards the *humanities*, based on their abilities and interests. While broad learning goals have been formulated nationally, there are no exams upon completion of lower secondary, which smoothens the transition to the upper levels but complicates the establishment of national attainment levels of these educational standards. Upper secondary education is concluded with national exams. More detailed information about the Dutch educational system, specifically the role of mathematics and physics, can be found in Tursucu (2019, pp. 24-26).

2.2.2 Scientific literacy

The compulsory part of science education is meant to effect 'scientific literacy' in students (Millar & Osborne, 1998; Ottevanger et al., 2014). Critical, scientifically literate citizens are capable of forming substantiated opinions on ethical and political dilemmas concerning science and technology (Aikenhead, 2005; European Commission, 1995). An important part of this literacy involves the ability to engage successfully in basic science inquiry and interpret scientific data and evidence, *the competency both to construct claims that are justified by data and to identify any flaws in the arguments of others* (OECD, 2013, p. 9). This competency is one of three core competences of scientific literacy and a major aim of science education (Next Generation Science Standards, 2013). It involves engaging faculties such as asking relevant questions, collecting and interpreting valid and reliable data, interpreting these data in acceptable ways, making informed choices based on these interpretations, and engaging in critical debate on each of these issues. Practical work in science is often expected to result in development of this competency.

2.2.3 Practical work

Practical work is expected to achieve more than developing aspects of scientific literacy or illustration of the empirical aspects of a science discipline. It is used to foster the understanding of scientific concepts, raise interest in scientific disciplines, teach practical skills (e.g., how to manipulate equipment), enhance students' ability to do science in which the aforementioned faculties are applied and teach students about the nature of science

(Abrahams, 2011; Dillon, 2008; Hofstein, 2017; Millar, 2010).

To attain these goals, secondary school students usually follow a prescribed, fixed procedure starting with a research question posed by the teacher. Students manipulate the given measuring equipment to collect the necessary data. They analyse and interpret these by answering scaffolding questions often provided in worksheets. As much of the work and thinking has already been done for the students in these highly teacher-directed activities they are called *closed* or *guided* (R. L. Bell, Smetana, & Binns, 2005; Tamir, 1991). However, it is demonstrated and often argued that these closed inquiries do not result in the understandings, attitudes and skills we want students to develop (Hodson, 1990, 1993, 2014; Hofstein & Kind, 2012; Holmes et al., 2017; Holmes & Wieman, 2018; Schwartz, Lederman, & Crawford, 2004; Wieman, 2015, 2016). Various authors recommend (gradually) more 'open' activities with more cognitive tasks for students to pursue inquiry learning (Banchi & Bell, 2008; Hodson, 2014; Hofstein & Kind, 2012; Holmes & Wieman, 2018; Wieman, 2015; Zion & Mendelovici, 2012). It is argued that students learn more and better when given opportunities to make decisions and to evaluate their decision (Hodson, 2014; Holmes & Wieman, 2018), since only then they are obliged to be engaged minds-on rather than merely hands-on (Abrahams & Millar, 2008; Hofstein & Kind, 2012). However, the limitations of the actual classroom often mean that students, in any type of practical, collect the data during class but analyse these and reach conclusions at home without the teacher's help so that the most demanding cognitive tasks (Wieman, 2015) are carried out without additional aid. Doing so successfully requires that they can perform an adequate, independent data analysis.

2.2.4 Data analysis in practical work

Previous studies of students' data analysis competence focused on students aged 14 and younger (Gott & Duggan, 1995; Gott & Roberts, 2008; Kanari & Millar, 2004; Lubben & Millar, 1996; Millar et al., 1994) involving only qualitative relationships ('more of A then more of B') or quantitative relationships at university level (Allie, Buffler, Campbell, & Lubben, 1998; Séré, Journeaux, & Larcher, 1993; Walsh, Quinn, Wieman, & Holmes, 2019). Other studies involving data-analysis focus on the students' knowledge of measurement uncertainties (Farmer, 2012; Kok et al., 2019; Stump, White, Passante, & Holmes, 2020) or the role and use of graphs (Lachmayer, Nerdel, & Prechtl, 2007; Pospiech et al., 2019; von Kotzebue, Gerstl, & Nerdel, 2015). Recent attempts to provide more coherence between mathematics and physics education (Boohan, 2016a, 2016b; Mooldijk & Sonneveld, 2010) show that the problem of using mathematics in physics, and data-analysis skills in particular is still pertinent and unsolved (Tursucu, 2019; Wong, 2018).

Compulsory science education is expected to prepare students for their pursuit of science at senior secondary level where they are expected to take agency in and organize independently their own research. It is important to establish a baseline of students'

competence in quantitative data analysis at this stage, since this freedom and autonomy can lead to successful performance only if accompanied by sufficient competence. Rather than passing judgement on how well students do, however, this study is meant to evaluate the outcomes of inquiry-oriented aspects of Dutch compulsory science education. We are unaware of other studies of these competencies in this age group and at this stage.

2.3 Theoretical framework

We present the *Procedural And Conceptual Knowledge in Science* (PACKS) as a model (Millar et al., 1994) to discuss the role of data analysis in practical work. We then use various curriculum documents to construct a consensus view on data analysis competences expected of 15 year olds. We conclude the theoretical framework with an overview and description of these competences.

2.3.1 PACKS

Since practical work often involves setting up and manipulating equipment and gathering data it may seem to mainly consist of hands-on activities. Millar *et al.* (1994), however, see practical work as a knowledge-based, primarily 'minds-on' activity where performance quality depends on access to pertinent knowledge. Their PACKS model (Figure 2.1) describes different types of knowledge that influence the choices students make in various stages of an investigation.

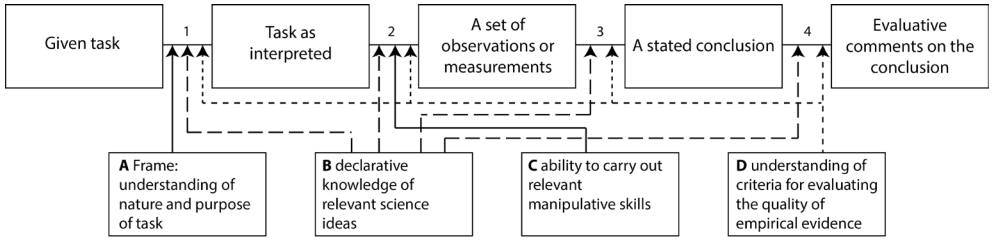


Figure 2.1: The PACKS-model (Millar et al, 1994) relates the various phases in inquiry and the types of knowledge affecting the decisions made in these phases.

While knowledge of type **A** mainly influences the initial stages in identifying the inquiry task, type **B,** understanding of the science relevant to the phenomenon, affects every step in the investigation. Type **C** includes knowledge involved in using appropriate equipment properly. The focus of this study is on knowledge of type **D***,* understanding the role of experimental data (Gott & Duggan, 1996; Millar et al., 1994). It includes an understanding of how measurement is subject to error, knowledge involved in reducing measurement uncertainties, and skills in assessing the reliability of data. It is indispensable in assessing the minimum number of measurements needed to establish a relationship, the

reliability of data, the (in)significance of differences in measured values, and drawing adequate conclusions (Millar, 1997; Millar et al., 1994). Millar (1997) therefore recommends practical work in which students themselves are enabled to assess the quality of experimental data and use a scientific approach to do so.

While curriculum documents often specify elements of type **D** knowledge in isolation, different types of PACKS tend to be applied in an integrated way (Walsh et al., 2019). We show below how we minimized interference between the types of knowledge in PACKS sufficiently to enable us to answer the research questions.

2.3.2 Concepts of evidence

Prominent elements in knowledge of type **D** are the *concepts of evidence* (CoE), including basic ideas such as repeatability, reproduction, precision and measurement uncertainty that underpin more abstract concepts such as reliability and validity (Gott & Duggan, 2007; Gott et al., 2003). The CoE guide decisions about how the practical work is set up, which measuring instruments are preferred, what types of patterns are detected, how anomalous data are treated, how a line of best fit is found to illustrate the underlying relationship, and how a defensible conclusion is derived from the available data (Gott & Roberts, 2008). Although not all CoE need to be understood in every scientific inquiry, understanding of the CoE supports the gathering of accurate experimental data and subsequently the drawing of acceptable conclusions supported by patterns or trends identified in the dataset.

2.3.3 Expected proficiency levels in data analysis

Over time, students should develop more understanding of evidence, and a more sophisticated understanding. So what level of understanding is required of students that have completed compulsory science education, what level of scientific literacy is satisfactory in this regard? We answer this question by constructing, below, an overview of CoE that need to be operationally understood by these students according to an apparent consensus among collected curriculum documents (Department for Education England, 2013; Jones, Wheeler, & Centurino, 2015; Ministry of Education Singapore, 2013; Next Generation Science Standards, 2013; Ottevanger et al., 2014; Spek & Rodenboog, 2011; United Kingdom Department for Education, 2014). We hold that this overview is feasible and sufficiently detailed for the purpose of this study. The abilities included in this study are printed in italics below and summed up in Table 2.1.

Students aged 15 are expected to be able to *spot trends*, *represent data graphically* and use statistical tools according to the OECD (2013, pp. 9 & 16). Scientific literacy includes the ability to account for the *uncertainty of measurements* (Ibid, pp. 8 & 16) and to assess whether *a claim is supported by data* (Ibid, pp. 9 & 16). These abilities accord with those specified in NGSS (2013, p. App. 57) for the ages of 12-14 in terms of *analysing and*

interpreting data and other curricula, e.g., the Science programmes of study: key stage 4 (United Kingdom Department for Education, 2014, pp. 5 - 6).

Dutch curriculum documents (Ottevanger *et al.*, 2014, pp. 18-20) largely paraphrase documentation of PISA (OECD, 2013) and the K-12 science education frameworks (National Research Council, 2000). The attainment levels, specifying data analysis abilities expected of 15 year-olds in The Netherlands, therefore align with these in other international curricula. For the Dutch school context, Spek and Rodenboog (2011) specify further requirements. According to these curriculum developers, a student should be able to:

(1) process data using a table and graph,
(2) draw a straight or curved line through a dataset while excluding erroneous data points, recognizing and estimating measurement errors,
(3) use the processed data to formulate one or more conclusions fitting the data, and
(4) compare the results and conclusions with an hypothesis.

This study explores whether students at age 15 can analyse and interpret experimental data adequately in terms of the aforementioned competences.

Table 2.1: Data analysis and interpretation skills for 15-year-olds in selected international curricula assessed in interview (I1-3) and practical (P1-3) probes.

Competence		**Probe number**						**# of docs**
Code	**Description**	**I1**	**I2**	**I3**	**P1**	**P2**	**P3**	
C1	Visualise data graphically and use/interpret these				v	v	v	56
C2	Establish the correct trend line.		v	v	v	v	v	109
C3	Qualitatively describe and identify the main features of a dataset or trend	v		v		v		71
C4	Describe qualitative similarities and differences between datasets					v	v	32
C5	Draw a conclusion which is supported by the dataset		v	v	v		v	79
C6	Justify the conclusions and indicate restrictions concerning the data analysis and conclusions			v	v		v	29
C7	Estimate the value of a variable using interpolation or extrapolation		v		v	v		66

In the absence of data from a nationwide exam or other means of testing, and in view of informal but frequent reports of teachers to the contrary, it is worthwhile to explore this baseline. While there is no formal (exam) program that specifies what data analysis competencies Dutch 15-year-olds should master, Table 2.1 summarizes seven competences as a tentative core derived from the relevant literature. These competences overlap within

the relevant curricula. The table comprises the criteria in this study for determining to what extent students have mastered data analysis and to describe the problems they encounter when applying it.

2.4 Method

2.4.1 Participants and setting

The study's setting is a modern, medium sized school preparing for higher vocational education or university studies. Information provided by the Inspectorate of Education (2018) regarding average final exam scores, number of students repeating years, and percentages of students electing a science stream shows that this school is situated close to or slightly above national averages in terms of participation and performance in science exam subjects. Hence these students can be regarded as representative of their age and ability group.

In terms of demographic characteristics as well, the 51 students (32 boys, 19 girls), with an average age of 15 and comprising two different classes, are not exceptional in that they are in majority autochthonous and generally from affluent families. They have opted for a science-based exam programme with physics as an exam subject, chosen by 30-40% of all students at the school.

A sample size of 51 accords with the average size of 40 samples in qualitative educational studies (Guetterman, 2015). It is both manageable and large enough to observe regularities and patterns as well as exceptional cases.

The lessons were taught by the first author who is the students' regular physics teacher. As a former trainer in an in-service professional development course focusing on practical work he is well aware of the challenges. In both action research (Altricher, Feldman, Posch, & Somekh, 2005; Carr & Kemmis, 2003) and educational design research (McKenney & Reeves, 2013; Van den Akker, Gravemeijer, McKenney, & Nieveen, 2006) teachers who research their own classroom practice are seen as situated close to the research-practice gap (Vanderlinde & Braak, 2010) and potentially equipped to close it (Bakx, Bakker, Koopman, & Beijaard, 2016). Potential threats to impartiality, objectivity and unconsciously influencing students (Trowler, 2011) were minimised by adhering to a pre-established interview protocol and virtually teacher-independent, worksheet-guided practical tasks.

From informal talks with teachers in our professional network, we infer that practical work and learning about inquiry receive above average attention at this school. The outcomes of the latest European PISA study suggest that our findings may be relevant to various other western European countries (Gurria, 2016).

2.4.2 Design

Two complementary approaches are used and triangulated to explore the competences listed in Table 2.1. The first uses interviews to explore what students think and do when asked to interpret data presented in a graph, and are prompted to provide as much information as they can. The second approach is used to study how students perform in a basic QPI task spontaneously, without assistance. In a qualitative, participatory research design (Bryman, 2015) detailed information is obtained about how students analyse and interpret data in three stages.

The first stage consists of three preparatory activities that develop in students the ability to *collect* sufficient and adequate data, making data pattern recognition possible. In stage 2, over the next three weeks, students are interviewed in pairs outside normal classes. Each interview consists of three probes, I1-3. In stage 3, three practical probes P1-3 are carried out. Each probe addresses several competences, though no single probe addresses them all: see Table 2.1.

Stage 1: Preparatory activities.

Preparatory activities are used to teach the necessity of repeating measurements, the importance of choosing an optimal data spread and range, and the purpose of averaging measured values. Students explore relationships between (1) body and arm length (Pols, Dekkers, & de Vries, 2019), (2) mass and period of a pendulum and (3), the distance travelled and number of cups propelled by a rolling marble (Farmer, 2012).

Stage 2: Interviews

In the interviews the researcher explains the purpose of the interview to the student pairs, asks for permission to make audio recordings while maintaining confidentiality and anonymity, and states that their answers do not influence their marks. He asks students to tell as much as they can, elaborate on what they think and express what they are looking at. Students then interpret the three graphs in Figure 2.2 by answering the questions in the Appendix. If only one student answers, the other is asked whether (s)he agrees or could add to the given answer, room is given for discussion when students are not in agreement.

Students are familiar with doing practical work and discussing their work, of any kind, in class with the teacher and each other. Although these discussions are normally not performed in interview style or recorded, there is no indication that this affected the validity or reliability of the data, or the content and form of what they brought forward.

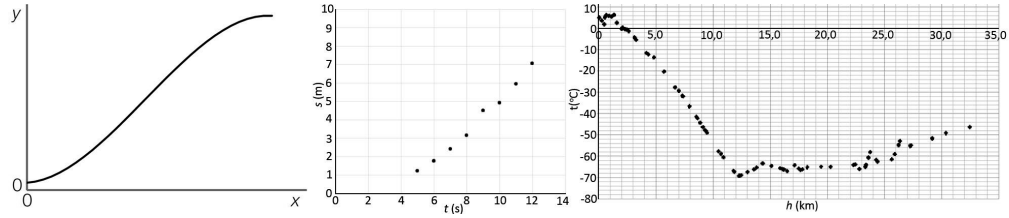


Figure 2.2: Graphs presented to interviewees in probes I1-3, respectively.

Probe I1. As in Boohan (2016b), students describe the graph on the left in Figure 2.2 to someone who cannot see it. This allows us to determine whether all essential parts of a graph are identified and described (C3, see Table 2.1.)

Probe I2. Presented with the computer generated data in a table and in the diagram central in Figure 2.2, students are asked, first, what the data might represent. This serves to determine whether students can interpret the observed data pattern (C1) in terms of any familiar event or episode in which quantities relate similarly. The graph is then said to represent the position-time graph of a marble, initially at rest, rolling down an incline. Students are invited to describe the pattern (C3), draw the trend line (C2), draw conclusions concerning the motion (C5 & C6), and predict two values (C7). Unless they note it spontaneously they are asked to consider whether the line of best fit includes the origin of the graph. Similarly they are asked whether some data points require special attention unless they note that the datum at t = 9 s is anomalous. Uniformly accelerated motion from rest is a topic that has been taught in the preceding two years.

Probe I3. Students are asked what the graph on the right of Figure 2.2 might represent. They are then told that it features temperature measurements of a weatherballoon by height. They are asked to identify and draw the trend line (C2), describe the pattern (C3), and explain whether they would provide this same description if they were writing a science report. Asked what conclusions might be drawn from the graph (C5), we expect them to identify two regions, one where the temperature decreases, and one where it increases. We do not expect them to identify, as an expert might, the troposphere, tropopause and stratosphere, respectively (Boeker & Van Grondelle, 2011).

The next question as to what additional data they would like to obtain to draw firmer conclusions, seeks to probe whether they can identify the limitations of their conclusions (C6). Students are finally asked which of the three probes they rate as the most difficult.

The audio recorded interviews of Probes I1-3 reveal what students are able to say about the issue if probed and encouraged by the interviewer who helps to bridge the 'knowing-doing gap' (Pfeffer & Sutton, 1999). The approach below is expected to yield less detailed answers, but does show what students do independently in a real class situation and thus has a higher degree of ecological validity (Brewer, 2000).

Stage 3: Practical Tasks

Students, working in pairs or an occasional triplet, carry out the three practical probes P1-3 (see Figure 2.3) guided by worksheet tasks that specify what should be measured. Further worksheet questions include: 'what is the shape of the graph?', 'what is the time required to travel X meters?', 'are there any outliers? if so, which?'. The worksheets were handed out at the start of the lessons. The answers to the worksheet questions together with audio recorded answers to teacher questions in P1 are the main data sources. The first two probes are conducted in regular physics lessons, the last by appointment with the technical assistant in the prep room, as is usual if only one setup is available.

Probe P1. After watching a movie scene of Spiderman swinging between tall buildings (Webb, 2012), where his motion can clearly be seen and the time for half a period can be measured accurately, student teams develop a plan to investigate whether this motion is realistic. Their plans are discussed in class to make sure they understand the task and know how to obtain highly accurate data. While data are gathered in pairs they are asked how they will go about establishing a relationship between variables from these. Thus is explored whether their actions are guided by a plan of approach, and whether they are aware of appropriate data analysis strategies.

The teams then predict the 'swing time' of a 5 meters long swing by using their measurements (C1, C2 & C7). Their predictions are tested in the next lesson by actual measurement. Using this additional data point, each team is asked again to predict Spiderman's swing time if the movie was realistic and thus to answer the research question (C5).

Probe P2. Teams establish the relation between the distance travelled and time required (C1 & C2) for a marble rolling on a horizontal track. As time is measured by hand, measurement uncertainty ought to be considered. Teams compare their datasets with those of others (C3 & C4), interpolate or extrapolate the travel time for two new distances (C7) and draw a conclusion about the type of motion of the marble (C5 & C6).

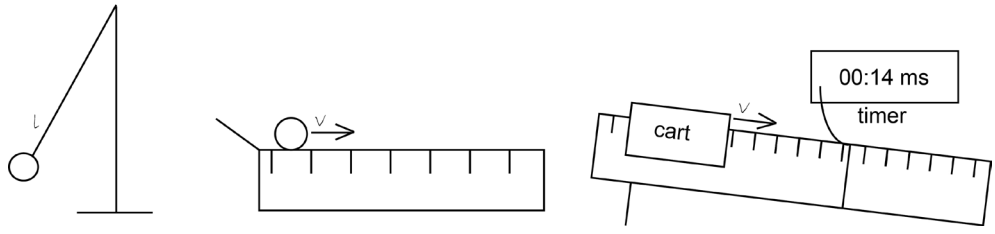


Figure 2.3: Practical probes P1-3 on analysis of data and drawing conclusions

Probe P3. Using a cart on a frictionless inclined track, the teams investigate the relation between travelled distance and time (C2) and compare their datasets with other pairs to investigate the influence of the incline. Measurement uncertainty plays almost no

role as an automated, accurate timer is used. Teams draw conclusions about the car's motion (C5 & C6) and predict its travelling times (C7).

2.4.3 Potential interference of other types of knowledge

A student lacking adequate knowledge of PACKS types **A**-**C** (of the research question, the situation at hand and the measuring techniques used to study it) may be prevented from succeeding in a probe even if the necessary type **D** knowledge is present, involving competence in analysing and interpreting the data. To exclude potential interference of this kind we used clear and simple inquiry tasks (**A**) employing familiar instruments and methods (**C**) in situations studied previously at a basic level (**B**). More specifically, probe I1 requires understanding only of the graph itself, how it came about is irrelevant to answering the questions. The situation of uniform motion in I2 ought to be familiar to students as it was taught previously, as will be verified. In I3, students are familiar with weather balloons and temperature measurements. Piloting with similar students shows that no interference is to be expected. In the interview probes students are invited to ask for clarification of aim, method or phenomenon before being questioned further.

The practical probes are introduced by the teacher or assistant demonstrating the equipment. Student responses are used to verify that students understand the purpose of the tasks, the nature of the situations and are familiar with the equipment, and thus enabled to answer the questions probing knowledge of type **D**. More specifically, in P1 we verified that students understood and accepted the model for Spiderman's motion, and that students are familiar with the equipment and measurement techniques (stopwatch, ruler, scales). A detailed knowledge of pendulum physics is not required here. In P2 we verified that all students were able to collect the required data, implying they understood what to do and how to use the instruments (stopwatch, rulers). Since students should be familiar with the kinematics of uniform motion we expect no interference of knowledge type **B** either. In P3, the teaching assistant helps students with their first measurement to familiarize them with the equipment. Students are expected to perform a simple comparison using the theory of uniform acceleration, taught in the preceding two years.

The interviews and practical probes are carefully monitored for signs of interference from other knowledge types and findings reported accordingly.

2.4.4 Data analysis

Following Schalk, Van der Schee, and Boersma (2008) who analysed students' application of CoE in biology, attainment criteria were specified for each competence and each probe. The three competence attainment levels (novice, intermediate and master) in Table 2.2 were defined on the basis of Lachmayer et al. (2007) for C1, (Boohan, 2016b) and (Lubben, Campbell, Buffler, & Allie, 2001) for C2; (Toulmin, 2003) and Gott and Duggan (2007) for C6. In all competences, the level 'novice' was allocated if none of the other levels applied.

Table 2.2. Level attainment criteria (where level 'Novice' is allocated if neither 'Master' nor 'Intermediate' apply).

Code	Competence	Master	Intermediate
C1	Visualise data graphically and use/interpret these.	In the graph drawn: (1) independent variable is assigned to x-axis, dependent to y-axis (2) axes are labelled (3) data points are plotted (4) scale is drawn, sensible range chosen.	One aspect missing
C2	Establish the correct trend line.	In the trend line established: (1) is considered whether the origin is part of the dataset, (2) is accounted for whether the data pattern is straight or curved, (3) the drawn line passes as closely as possible to the data points	(1) or (2) ignored
C3	Qualitatively describe and identify the main features of a dataset or trend.	Description includes I1: labels, starting point, 3 features of shape I3: labels, 3 regions P2: graph's shape, technical specification ('gradient', 'slope'), essential features clear enough to reproduce the graph	No more than two missing
C4	Describe qualitative similarities and differences between datasets.	A correct qualitative comparison of shapes and gradients of graphs of different data sets is given	One comparison missing or incorrect
C5	Draw a conclusion which is supported by the dataset.	All of the following are satisfied: (1) the conclusion answers the research question, (2) the conclusions is backed by the data, (3) the most extensive conclusion is drawn that fits the data	Only (1) and (2) are satisfied
C6	Justify the conclusions and indicate restrictions concerning the data analysis and conclusions.	Conclusions are actively supported though explicit backings, qualifiers, rebuttals and warrants. One at most is missing of: (1) provides the reasoning through which the data support the claim; (2) accounts for level of data accuracy; (3) discusses limitations of data, analysis, claims (4) accounts for assumptions underlying conclusions	(1) and (2) or (3) are present.
C7	Estimate the value of a variable using interpolation or extrapolation.	(1) Predicted values accord with theoretical or teacher's values. (2) Justification is given of the method.	Suitable method is found but not well applied.

For each probe and each pair, the attained competence level was determined. Since each probe addresses only a subset of competences the number of relevant probes per competence varies (see Table 2.1). For each competence and level, the number of probes in which that level is attained was divided by the total number of probes addressing that competence. These fractions are interpreted as the competence levels of the whole sample. E.g. since 'mastery' of competence C4 was observed in 32 probes out of 109 probes relevant

to C4, we express that level as: *Master C4 = 32/109 = .29*. While these numbers provide an overall, global description of the attainment levels of the sample, we provide a more detailed, qualitative illustration of each competence using students' thoughts and actions as expressed in the documents.

The analysis of attainment levels was carried out independently by a second researcher for 20% of the probes, randomly selected from the total sample. No significant differences with the original analysis were found, we regard the initial analysis as valid and reliable.

2.5 Results

An overall description of the attainment levels across different competences, constructed as outlined above, is given in Figure 2.4. Immediately obvious is that master level is attained across the sample only in competence C1. The figure also shows that these students generally do not make use of argumentation to justify their conclusions (C6). A further, qualitative elaboration of the data per competence is presented below. Labels such as 'P2-10' are used to refer to, in this case, the record of student pair 10 performing probe P2.

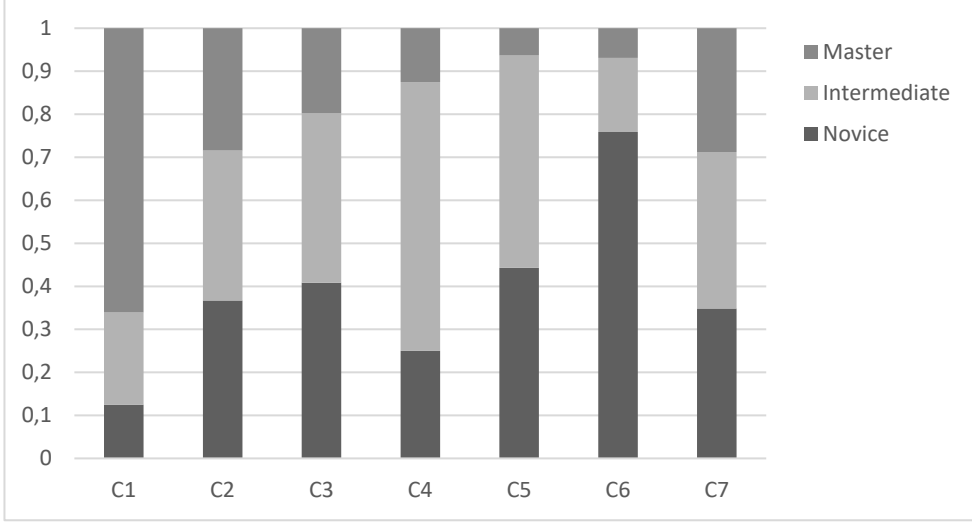


Figure 2.4: Attainment levels per student pair in each competence (N=51).

2.5.1 Exclusion of interference of PACKS types A-C

Since in each of practical probes P1-3, all responses include an appropriate graph of relevant data and attempts to identify pertinent relationships between one or more salient dependent and independent variables, students' PACKS of types **A**-**C** appears adequate for

these tasks. In interview probes I1-3 (see the Appendix), questions Q1-4 require answers about the appearance of the graph that do not require PACKS types **A**-**C**. Questions Q5-7 require PACKS of types **A** and **B** and findings are discussed among the relevant competences, C5-7.

2.5.2 Attainment level of competences

Competence 1: Visualising Data All students construct suitable graphs to visualize the data, including labels, units and suitable scales. Figure 2.5 displays a typical and correct graph. Aside from some students who do not place the independent variable on the horizontal axis (14%) or do not label the axes at all (12%), most graphs (67%) meet all scientific conventions.

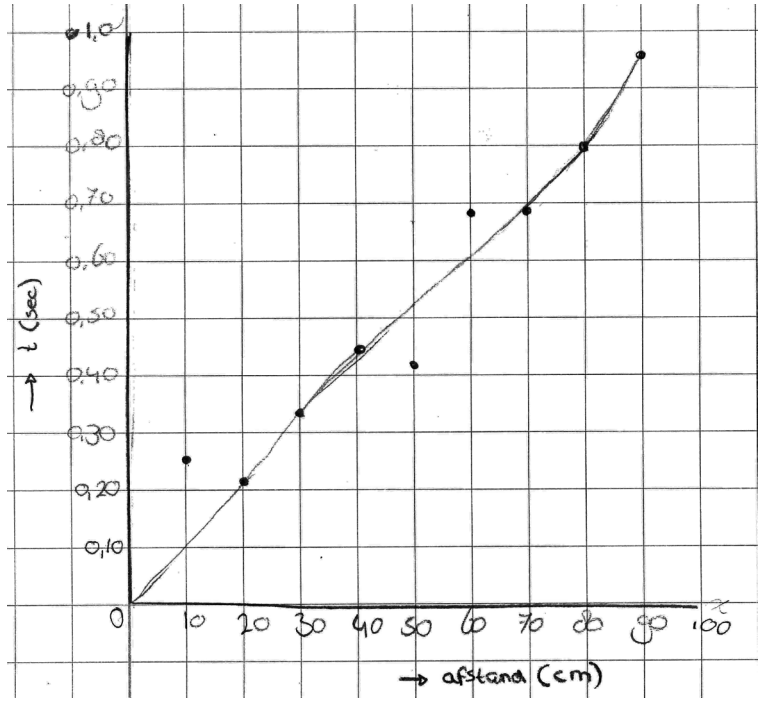


Figure 2.5: A representative graph from P2 (P2-3) with travelled distance displayed on the x-axis and the measured time on the y-axis. Scientific conventions (C1) are satisfied, the trend line is smooth, partly connecting the data points (C2).

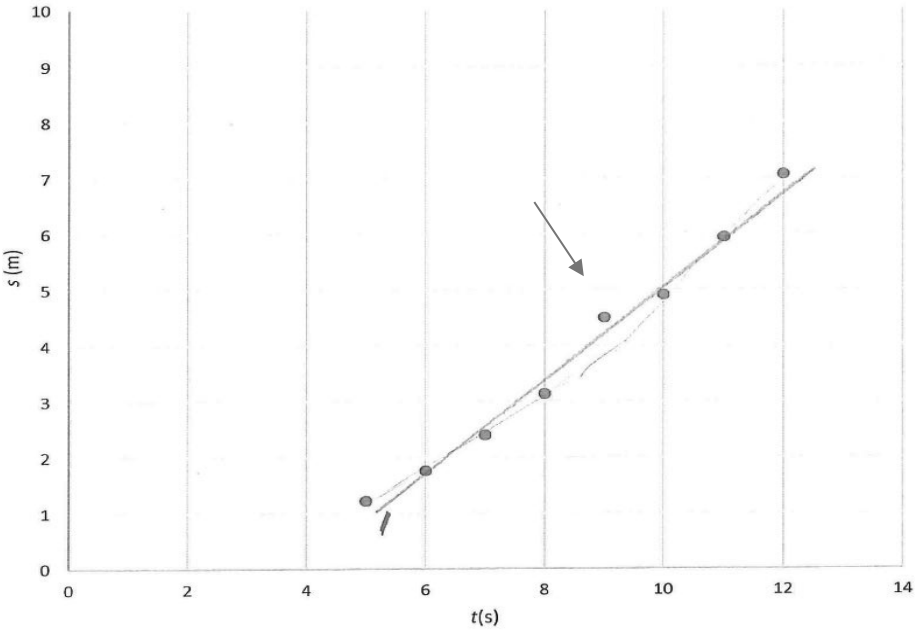


Figure 2.6. A straight line is drawn in I2, disregarding the identified anomaly at t = 9 s. As the graph displays the motion of a ball, starting from rest, in a travelled distance versus time graph, the origin should be used as well.

Competence 2: Drawing a Trend Line Students' ability to draw a trend line differs substantially across probes, as is illustrated in Figures 2.5-2.7. Some students simply connect the dots, or disregard measurement uncertainties in some other way (41%). Students often cannot produce a trend line, either because they do not have the concept or are unable to construct it. If present at all, the concept of a trend line is not expressed with detail or precision, as is illustrated in this dialogue between the researcher (R) and a student (S) during probe P1:

R: How will you establish the relationship between the variables?
S: Draw a graph. (R: And then?) Looking whether it is a linear relationship, or another odd one… (R: By?) By looking how the line goes.
R: And if is not a linear relationship, then?
S: Then it is a different kind of relationship. Or none at all.

Similarly, many students look for a linear trend but are unaware of alternatives if the pattern is different, as is reflected in their graphs in P1. Yet, if they obtain accurate measurements in P1, many students (60%) draw the curved line of best fit suggested by the data. However when one datum is added to the dataset well beyond the initial range for a very long swing, about half the students revert in some way to drawing a straight line, as shown in Figure 2.7.

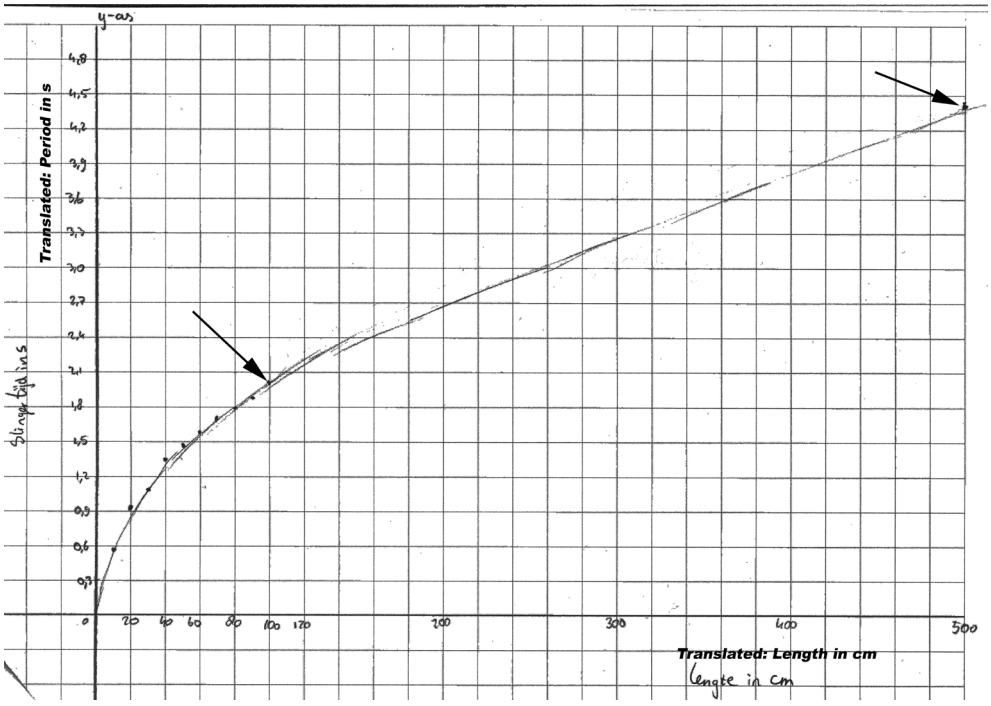


Figure 2.7. In probe P1, students investigate how the swing time (y-axis) is affected by the length of the swing (x-axis). Some students combine a curved and a straight line to connect the last datum of initial dataset with the additional datum at l = 5,0 m, both indicated with arrows.

In I2, the fact that the movement starts from rest implies that the graph must start horizontally and curve upward. 50% of the students do not provide this interpretation, but rather than failing to interpret the given they appear to ignore it altogether.

Competence 3: Describing a Dataset In the interviews students experience exceptional difficulties in describing a graph verbally. Important features such as, e.g., the process it represents, the value of the y-intercept or the shape are omitted. Most students fail to use (correctly) such scientific terms. The information density of the graph seems to influence the quality of descriptions negatively, as graph I3 tends to be described less elaborately than I1. The students' struggle is illustrated in the following (I1-25):

R: How would you describe the graph to someone who does not see the graph?
S1: Upward with curves.
S2: Increasing rise, decreasing decline.
R: Have you described all essential features?
S1: I expect it to stop at the end.
R: What do you mean with stop at the end?
S1: That it runs horizontally. [No further details are given.]

This excerpt contains a mixture of scientific terms and colloquial terms. As in many interviews, the description is brief and qualitative, it does not provide insights in the most essential features of the graph. Some students' statements are categorized at 'master' level, but even those tend to be brief. A few statements, such as the following quote from P2-10 provide an extensive, correct description of a fair number of characteristic but also lack some essential information (e.g. its slope):

> 'On the x-axis is the distance in cm with a 10 cm interval. The time is on the y-axis with a 0.1 s interval. The line is fairly linear. We have also drawn the maximum, minimum and average value.'

Students (60%) report at the end of the interview that describing the graphs is hard and that they are unsure what to describe. Several students confide that in a practical report, they would provide more information and use more difficult words as that would probably lead to a higher mark as it would be more likely to include the required details.

Competence 4: Comparing Datasets Over 60% of students' answers are categorized as intermediate because they include either a comparison of shape or steepness, but rarely of both. Students compare the datasets merely qualitatively and by sight, superficially. Comparisons are brief and provide little insight into the precise differences and similarities, even if students were specifically asked to do so. The tools for comparing datasets that students ought to have acquired in mathematics are not applied at all. The following are typical examples at intermediate and master level, respectively (P2-8; P3-22):

> 'Same relation. A few deviating points. Our line goes through most points. Theirs through all.'
> 'We both have a quadratic relation. Their graph is slightly steeper than ours.'

Competence 5: Drawing Conclusions Students' answers reflect an understanding of the situations and the questions asked about it (PACKS types **A** and **B**), in accord with prior learning and our expectations. In respect of C5, a conclusion is rated 'intermediate' if it answers the research question and accords with the data – an optimally informative answer is not required. The substantial 'intermediate' fraction in C5 (Figure 2.5) therefore includes superficial conclusions such as, e.g.:

> 'as time increases the travelled distance of the marble increases' (25% in I2),
> 'the marble accelerates' (50% in I2).

Many of the conclusions simply restate the results, e.g. for I3-28:

> 'The higher you go, the colder it becomes. The temperature doesn't fall below -70 above 12 km. At 2000 m it drops below 0.'

Competence 6: Specifying Limitations The first two examples in the above section illustrate that students often do not clearly link the data to their conclusions. The specific evidence supporting a particular claim, or whether that evidence suffices to make the claim acceptable is not provided. There are exceptions to these superficial, unsubstantiated conclusions, but only in P1 where most students conclude that Spiderman's swing in the film is unrealistic, e.g. (P1-2&5):

> 'the swing time is much smaller than we found using the pendulum'.
> '**If our calculations are correct** and **the sling is indeed 50 m long**, then the motion of Spiderman in the movie is not possible.'

Two groups include qualifiers (with an example printed above in bold) in their conclusion. However, students generally do not support their claims with arguments.

Since students answers to I3-Q5 & 6 reveal an insufficient knowledge of the atmosphere (PACKS type **B**) to suggest appropriate ways to address the limitations of the given measurements, these data are excluded from further consideration here.

Competence 7: Predicting the Values of Variables Students are generally aware of extrapolation and interpolation as techniques for estimating values based on measurement, and show adequate PACKS types **A**-**B** in all probes. In cases where the mathematical relationship between the two physical quantities is directly proportional and the measurement uncertainty small, they apply these techniques correctly. However if the uncertainty in the measurements is more substantial, students tend to incorrectly connect the data points (C2) and fail to interpolate properly. Some students use a direct proportional relation when they identify it as a linear relation, forgetting to take the y-intercept into account. In P1, half of the students extrapolate values based on a direct proportional relationship though the trend is curved. Two groups correctly predict the 'swing time' of a 5 m long swing but fail to explain why their method is justified.

2.6 Discussion

Our findings accord with previous studies reporting that students of various ages have difficulties in analysing data (Bailey & Millar, 1996; Kanari & Millar, 2004; Lubben & Millar, 1996; Millar et al., 1994). In analysing the data, we confirmed the students' ability to visualize the data graphically and look for a linear pattern. Although students learned in mathematics how to establish proportional, linear and squared relations and studied various aspects of data analysis, they have no strategy available to analyse the data unless the relationship is directly proportional. The failure to apply in physics what was learned in, e.g., mathematics is known as a transfer problem (Boohan, 2016b; Leinhardt, Zaslavsky, & Stein, 1990; Wong, 2017). This study extends results reported there about qualitative data

analysis to the students' approach and the difficulties experienced in quantitative data analysis and, below, in drawing conclusions based on empirical data.

Throughout all tasks, students make mistakes and encounter difficulties that prevent a successful data analysis and lead to merely superficial conclusions. Students tend to ignore or be unable to use all relevant information in constructing a graph. They have trouble in describing graphs and are not fully aware of their purposes in science reports. They do not always distinguish linear relationships from other types, but if they do, fail to apply that insight correctly in predicting a value. When analysing the data, they hardly use their existing knowledge of physics or mathematics. As a consequence, their conclusions are often superficial, unsubstantiated and without specification of limitations to their validity or reliability. Out of seven higher order data analysis competences expected of students at this level, only one, drawing a graph given a quantitative dataset, is attained at an acceptable level in the sample.

The participants in this study enjoyed some of the best compulsory science education available in the Netherlands, belonged to the top 40% of Dutch students in terms of academic ability and had, since they had elected a science-based exam programme, expressed an interest and willingness to learn science. Our results indicate, albeit in a small sample but without any reason to regard these students as special, that in the area of data analysis and interpretation they have not attained scientific literacy as specified by national and international curricula, nor the level assumed at the start of post-compulsory science education. Addressing the problem seems to be relevant.

We tentatively identify four areas of concern in mapping a way forward:

(1) Students rarely attach intrinsic (scientific) relevance or value to the questions or problems we present (Hodson, 1990, 2014). They are quite willing to please their teacher and do as they are asked but the issues at hand seem rarely to relate to any concern or interest of their own. They are quite satisfied with a common sense or superficial answer to the questions, but note that in a report they would embellish their account by using 'difficult' words in the expectation of obtaining a higher mark. We should try to engage students in inquiry that they too see as relevant and worthwhile if we expect them to invest in learning how to do inquiry well.

(2) As found by others (Abrahams & Millar, 2008; Millar et al., 1999; Pols, 2020a), students happily leave judgement on the quality of the answer to the teacher or another external authority. This is arguably not conducive to an approach that takes on the rigour and thoroughness required in scientific inquiry. The Spiderman probe (P1) provides a notable exception: here, some students clearly *are* personally interested in finding out if movies depict Spiderman's movements realistically, and eager to carry out additional

measurements on their own to obtain more information about what happens if a swing's length increases from one to five meters. Students engaged in personally relevant and worthwhile inquiry can be expected to become interested in finding useful and trustworthy answers, and stimulated to take responsibility for finding these.

(3) Useful, trustworthy answers are optimally supported by cogent arguments connecting claims to data. Inquiry ought to aim at students constructing these arguments and develop the understanding that a scientific approach is optimally suited to doing so (Gott & Duggan, 2007; Hodson, 2014; Hofstein & Kind, 2012; Woolgar & Latour, 1986).

(4) Students need to develop an understanding of what *counts* as convincing evidence in a scientific argument, and develop an understanding of the CoE (Gott & Duggan, 1996, 2007; Lubben & Millar, 1996; Millar, 1997; R. Roberts & Reading, 2015; Wellington, 2002).

This combination of characteristics is probably necessary, but not necessarily sufficient in constructing a viable pathway of inquiry learning. Our current research is directed at developing and evaluating these ideas in a practical sense (Pols et al., 2019). We think we may do so by changing students' task perception to that of an experienced scientific researcher: find and defend the *best possible* answer given the circumstances. This study provides both instruments that are useful in establishing what learning takes place, and a starting point for deriving suitable learning activities.

The findings imply students should develop their data-analysis skills before they can be expected to apply these independently in QPI. To this end, the probes can be adapted. Probes I1-3 can be carried out using a classroom discussion. As Wellington (2002) suggests: after students elaborate their views, the teacher discusses what features are essential, what patterns can be detected, and if available, how theory of the phenomenon might help in doing the analysis. In this way, *the students' conclusions are valued, discussed and related to the teacher's hoped-for conclusion* (Tasker & Freyberg, 1985).

Data gathered in the practical probes can be shared on the interactive whiteboard and discussed with a focus on similarities and differences in the datasets. Sharing the measurements reveals spread in measurements, similarities in shape, difference in slope etc. In this way, the central part of practical work is the discussion and meaning making of the data rather than merely gathering data to confirm a known relationship (Abrahams & Millar, 2008; Gunstone & Champagne, 1990).

2.6.1 Limitations

Eventually, we will be interested in establishing whether secondary students can develop and apply in an integrated way the full scope of PACKS. This study addresses only a small

section of it, referred to as PACKS type D, where potential interference of types A-C was reduced as far as possible. Therefore, while a baseline of this knowledge type is established for a specific age and ability group, a more comprehensive approach will eventually be required. Similarly, we used a small scale, qualitative approach, providing our baseline with depth and detail. While we have argued that the findings have a fair degree of generalizability, a more quantitative and large scale confirmation may be desirable.

2.7 Conclusions

This study investigates Dutch students' ability to analyse and interpret experimental data by constructing adequate data representations and drawing qualified, appropriate, defensible conclusions from these data. Seven associated competences, distilled from contemporary curricula in the international literature, specify what we can expect in that area of 15-year-olds in a science-based program. The attainment level in these competences is established in a sample of 51 students of two intact science classes using three thinking-aloud and three practical probes. Both approaches show that students encounter many difficulties in analysing and interpreting data. The competence of constructing a graph from given data is the only one that was adequately mastered. We argue that *neither* the students' academic level *nor* the quality of the education they had received can account for these results. There is every reason to expect that these results are relevant in much wider settings than the one studied here. In this particular respect, the outcomes of Dutch compulsory science education fall well short of realising scientific literacy for all. The findings further imply that before students can benefit from more 'open' practical work in developing independent scientific research competences, they first or concurrently need support in developing the data analysis and interpretation competences studied here. Before students can be expected to analyse and interpret data without help, they will have to overcome the many problems uncovered or confirmed here.

2.8 Reflection and next step(s)

This first study showed that students indeed exhibit an insufficient ability to analyse empirical data so to engage in practical work independently. However, it also yielded a more worrying result: students are already hindered in an earlier stage which prevents them from successful engagement in practical work. They lack a sense of scientific purpose: to find and defend the best answer obtainable in the given circumstances (areas of concern 1 & 2). As a result of these findings, the study's focus shifted: from practical work in general with a focus on data-analysis towards *learning to engage in scientific inquiry*. In the next chapter we address areas of concern 3 & 4, by specifying what learning to engage in scientific inquiry entails with a focus on convincing evidence.

3. Defining and Assessing Understandings of Evidence with the Assessment Rubric for Physics Inquiry - Towards Integration of Argumentation and Inquiry

Article previously published as:

Pols, C.F.J., Dekkers, P.J.J.M., and de Vries, M.J. (2022). Defining and Assessing Understandings of Evidence with ARPI: Towards Integration of Argumentation and Inquiry. *Physical Review: Physics Education Research,* DOI: 10.1103/PhysRevPhysEducRes.18.010111

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 engage in 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 study we present a view of inquiry knowledge and a way to assess it that acknowledges both of these components. We deconstruct 'inquiry knowledge' as a set of 'Understandings of Evidence' (UoE) - insights and views that an experimental researcher relies on in constructing and evaluating scientific evidence. Acquirement of these insights 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. 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 (ARPI), is validated in this study. Preliminary results suggesting a high degree of ecological validity are presented and evaluated.

3.1 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 (Hodson, 2014). 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 (Millar, 2004; Millar et al., 1999). In teaching students how to do science and supporting them in developing the required knowledge, it is helpful to see an inquiry as the building of a scientifically cogent argument (Driver et al., 2000; Gott & Duggan, 2007; Hofstein & Kind, 2012). 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 (Gott & Duggan, 2007; Toulmin, 2003; Woolgar & Latour, 1986). 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 (Barron et al., 1998; Black & Wiliam, 2005; Hodson, 1992; Walsh et al., 2019). However, tools for measuring student's understanding of experimental physics are scarce (Holmes & Wieman, 2018). Frequently used 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 (Lewandowski, 2014; Walsh et al., 2019; Zwickl, Hirokawa, Finkelstein, & Lewandowski, 2014). The Scientific Abilities Assessment (Etkina et al., 2006), 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 (Giddings, Hofstein, & Lunetta, 1991; Hodson, 2014; Holmes & Wieman, 2018). There remains a need for standardized, objective assessment criteria and instruments to assess the degree to which students develop inquiry understandings and skills (Holmes & Wieman, 2018; Walsh et al., 2019). In this study, we construct an approach to derive students' grasp and use of the proposed understandings from the substantiations and justifications of choices they make and actions they carry out during inquiry.

3.2 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) (Millar et al., 1994). 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.

3.2.1 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. (1994) 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 recognised 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 (Pols, Dekkers, & de Vries, 2021).

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* (Kelly, 2014, p. 329), 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 (Cavagnetto, 2010; Duschl & Osborne, 2002; Erduran & Jiménez-Aleixandre, 2008; Erduran et al., 2005; Erduran et al., 2004; Hofstein & Kind, 2012). 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, analysing, judging, criticising 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* (Gott & Duggan, 2007).

3.2.2 PACKS and the Concepts of Evidence

In their PACKS model presented in Figure 3.1, Millar et al. (1994) 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.

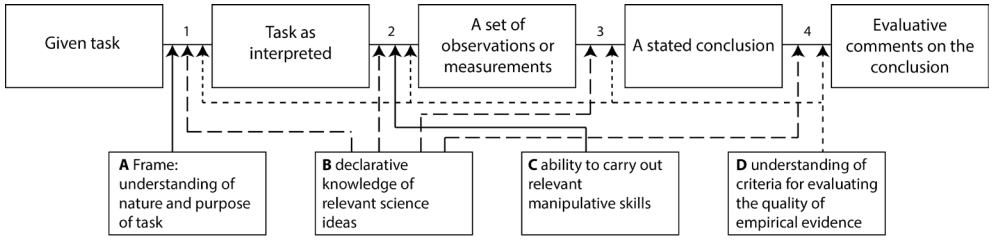


Figure 3.1: In their PACKS model, Millar et al. (1994) link how the taking of decisions during various stages of an inquiry are informed by different types of knowledge.

In their study, Millar et al. (1994) 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' (Gott & Duggan, 1996, 2003; Gott et al., 2003). 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 (Gott et al., 2003). Gott et al. (2003) 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 (Gott & Duggan, 2007; Gott & Roberts, 2008; OECD, 2013; Osborne, 2014b; R. Roberts & Johnson, 2015; R. Roberts & Reading, 2015).

However, there is a complication in assessing students' understanding of each separate CoE. Individual concepts acquire their meaning through a network of interrelated concepts (White & Gunstone, 1992), 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.

3.3 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 minimise 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 (Lachmayer et al., 2007), 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 (Gott et al., 2003). 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 that **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 recognise 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 operationalise 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.

3.4 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 inquiry in physics. We consider the UoE required in QPI: 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?*

3.5 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).

3.5.1 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 (Cohen, Manion, & Morrison, 2013, pp. 256-257). 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 (Allen & Knight, 2009; Kempa, 1986). 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 (Hsu & Sandford, 2007), 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 (Delbecq, Van de Ven, & Gustafson, 1975). Experts' input can be obtained by questionnaires or other means of data-collection (Murry Jr & Hammons, 1995). In a modified Delphi technique experts are presented carefully selected items stemming from e.g. a literature study (Custer, Scarcella, & Stewart, 1999; Murry Jr & Hammons, 1995) 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 (Hsu & Sandford, 2007). The required number of iterations depends on how quickly experts' views converge. While often three iterations suffice (Custer et al., 1999; Hsu & Sandford, 2007), sufficient convergence in this study (Table 3.1) was attained in round 4, after two. Iterations 1 and 2 of the Delphi section of the study 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 (Hsu & Sandford, 2007), it is convenient to describe the different kinds of expert participants alongside the successive rounds of the design.

Table 3.1: The participants in the modified and augmented Delphi study and the research rounds they were involved in.

Participants	#	round 1 Literature review	round 2 Delphi Iteration 1	round 3 Ecological validity Field test	round 4 Delphi Iteration 2	round 5 Ecological validity Expert interviews
Content Experts	8		questionnaire		interview	interview
Experts of Practice	5			interview		
External Experts	6					interview

3.5.2 Participants, instruments and analysis

First round: Prototype on the basis of personal professional expertise & 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 & 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 (1996) as it potentially reduces the number of required iterations. We expected the instrument to have acquired *face validity*.

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 (Cohen et al., 2013, p. 219). 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 3.2, agreed to participate. The sample size is well within the range of three to ten recommended by Rubio, Berg-Weger, Tebb, Lee, and Rauch (2003).

*Table 3.2: Description of the eight participating content experts in terms of five expert criteria. The symbols **p**, **c** and **i** denote that the criterion was satisfied in the **p**ast, **c**urrently or **i**n progress, respectively. Symbols **s** and **u** denote **s**econdary school level and **u**niversity level, and **d** denotes a Doctorate in **ph**ysics or in physics **ed**ucation*

Expert	Physics teacher	University Lab course teacher	Physics teacher trainer	PhD
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			

Instrument & 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 analysed 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.

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 3.2) 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 & 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 *semistructured 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 experienced 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.

Fourth round: Delphi iteration 2 - Determining the consensus between content experts

Goal: The final 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 & 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 (2006) 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.

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. Semi-structured, 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 3.1). This group of PhD's in physics were found by means of convenience sampling from a Faculty Online Learning Community (FOLC) (Dancy, Lau, Rundquist, & Henderson, 2019). Their involvement contributes to the external validity of ARPI (Gast, 2014; Kratochwill, 2013) 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 et al., 2013, pp. 674-685).

3.5.3 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.

3.6 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.

3.6.1 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 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 (Johnson, Britain, & Unit, 1989; Millar et al., 1994; Welford, Harlen, & Schofield, 1985), and Kempa's model of doing science which is recognized to be useful for assessment (Hodson, 1992; Kempa, 1986). As shown in table 3.3, the phases of ARPI integrate the phases of the other two models.

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 (European Commission, 1995; Millar, 2008; Millar & Osborne, 1998; NRC, 2013; OECD, 2013; D. A. Roberts & Bybee, 2014). 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 (Harrie Eijkelhof, 2014; Netherlands Institute for Curriculum Development, 2016; Ottevanger et al., 2014). Other international curricula in the English-speaking world are similarly derived from these sources (Breakspear, 2012; Burdett & Sturman, 2013; Ministry of Education Singapore, 2013; Singapore, 2019; Sunder, 2016). Therefore, we consider the OECD (2013) and NRC (2013) documents to be adequate and sufficient in their description of the learning goals for secondary school level physics inquiry.

Table 3.3: The phases of ARPI overlap with the phases as distinguished in the APU model and by Kempa.

<table>
<tr><th>Kempa</th><th>APU</th><th>ARPI</th></tr>
<tr><td>Recognition and formulation of the problem</td><td>Problem formulation</td><td>Asking questions</td></tr>
<tr><td rowspan="2">Design and planning of experimental procedure</td><td>Planning an experiment</td><td>Design</td></tr>
<tr><td rowspan="2">Carrying out an experiment</td><td rowspan="3">Methods & Procedure</td></tr>
<tr><td>Setting-up and execution of experimental work (manipulation)</td></tr>
<tr><td>Observational and measuring skills (including the recording of data and observations)</td><td>Recording data</td></tr>
<tr><td rowspan="2">Interpretation and evaluation of experimental data and observations</td><td>Interpreting data & drawing conclusions</td><td>Analysis</td></tr>
<tr><td>Evaluation of results</td><td>Conclusion & Evaluation</td></tr>
</table>

At the tertiary level, physics education aims at teaching students to *think like a physicist* (Kozminski et al., 2014; Redish & Rigden, 1998; Van Heuvelen, 1991). Wieman (2015), 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* (Kozminski et al., 2014). Furthermore, the frequently referenced source Etkina et al. (2006) 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

3.4 (NRC, 2013, pp. 48-66). 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.

Table 3.4: 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

On the other hand, the initial list did not include the provision and reception of feedback although, as Driver et al. (2000, p. 288) 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 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 (Brookhart, 1999; Moskal, 2000; Rusman & Dirkx, 2017) but found no consensus (Rusman & Dirkx, 2017). Moskal (2000) and (Brookhart, 1999) 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.

3.6.2 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 3.5 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.

Table 3.5: 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' (J. S. Lederman et al., 2014) 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 (Hodson, 2014).
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 an 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 & communication.	Clarification of content & aims.	Issues pertaining to presentation and communication are assessed, but as integral parts of the expression of UoE's.
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 & consistency of text.

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.

3.6.3 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 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.*

3.6.4 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 analyse 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 analyse 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 3.5, 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 3.6, 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*.

Table 3.6: 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.

	#	The researcher understands that:	This understanding is demonstrated by:	Highest level	Intermediate level	Lowest level
Phase 1 Asking questions	1	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
Phase 2 Design	4	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
Phase 3 Method & Procedure	8	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
Phase 4 Analysis	11	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.

	#	The researcher understands that:	This understanding is demonstrated by:	Highest level	Intermediate level	Lowest level
Phase 5 Conclusion & evaluation	14	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 insuf-ficiently infor-mative, 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 trust-worthiness of the data and the outcomes in the given circumstan-ces 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
Phase 6 (Peer) Review	19	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.

3.6.5 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.a., 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.

One external expert, a member of the *AAPT Committee on Laboratories* providing recommendations for the undergraduate physics laboratory curriculum (Kozminski et al., 2014), 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 (Biggs, 1996) is indispensable and we hold that ARPI, or the underlying ideas on which the construct is based, is functional in maintaining it.

3.7 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.

3.8 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 (1997) 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 (Millar et al., 1994) on the basis of the understandings specified in ARPI.

3.8.1 Limitations and future research

ARPI was constructed with a focus on knowledge type **D** in the PACKS model (Millar et al., 1994) by organising interrelated CoE (Gott et al., 2003) 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 (Walsh et al., 2019). 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 (Dounas-Frazer & Lewandowski, 2018; Zwickl, Hu, Finkelstein, & Lewandowski, 2015) - 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-learning 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.

3.9 Reflection and next step(s)

Now that we have specified what learning to engage in QPI entails – have identified the learning goals – we can build activities that target these learning goals, i.e. the UoE. Ideally these are not isolated events (Dekkers, 1997) but are part of a coherent structure where the activities build upon each other and in which students can relate the activity to previous ones. In other words, the activities should be part of a teaching-learning sequence (TLS). In the next chapter we describe the TLS that has been designed, where we present details pertaining the activities that are especially relevant for teachers.

4. Introducing argumentation in inquiry – a combination of five exemplary activities

Article previously published as:

Pols, C.F.J., Dekkers, P.J.J.M., and de Vries, M.J. (2019). Introducing argumentation in inquiry – a combination of five exemplary activities. *Physics Education,* DOI: 10.1088/1361-6552/ab2ae5

Successfully carrying out a secondary school physics inquiry requires a considerable amount of procedural and content knowledge. It further requires knowledge of how and why maintaining scientific standards produces the best available answer to the given research question. To this purpose, a series of five inquiry activities was developed and tested in a single case study with students aged 14. The test shows that students indeed come to use a more scientific approach to inquiry tasks and understand why they should do so. We believe that this series of activities can serve as a starting point for more complex physics inquiries.

4.1 The problem of teaching inquiry skills

Inexperienced students often use inadequate procedures in scientific inquiry of, e.g., the pendulum. They frequently choose only two values for the length instead of a wide range, measure only once at each length instead of repeating and calculating averages, and draw a straight line through the data-pattern that (to us) clearly looks curved (Gott & Duggan, 1995). Textbooks often 'help' students so they merely have to fill in a table as instructed, calculate averages and square roots, and plot a graph that is *meant* to be straight. This often precludes their exploration of further assumptions about the pendulum, and many remain mystified as to why the square root was taken. Worse, however, is that if these issues are not addressed at an early stage they will re-emerge years later and cause further problems. Yet, explaining *why* procedures should be followed rarely helps. While students tend to comply and do as they are told, they stop doing so when we stop telling them to. Could it be that they fail to see the point of doing so if all we ask is: how does the period of a pendulum depend on its length? Can we expand the students' aim from answering the research question to finding the *best possible* answer, and *demonstrating* that it is? We present a series of five activities designed for this purpose and our experiences in a class of 21 students aged 14.

4.2 The activities of the teaching-learning sequence

4.2.1 Activity 1 – Investigating what they know in the Pirates' Pendulum

During the making of a pirate film Captain Jack Sparrow and his mates are spectacularly swinging between ships of war, explosions going off and razor-sharp weapons flashing everywhere. Students need no convincing that the stunt coordinator must have a thorough understanding of the swinging, since Jack should arrive at a given spot immediately after the explosion, not during.

Students explore the physics of a pendulum to provide the stunt coordinator with the required information. Students identify factors they think influence the swing time and investigate these in small groups. The teacher monitors, asking supporting questions with the final discussion in mind:

- Can you explain what you are doing there? Why? What are you trying to find out?
- How do you carry out your measurements? What instruments do you use, and why?
- What will you report? Why should the stunt coordinator trust your results?
- What could you do to make your results even more trustworthy?

Students' actions and conclusions are as usual, a report to the fictitious coordinator is the only new element. Rather than on the findings, however, the final discussion focuses on the question: if you were a stunt(wo)man, knowing what information the stunt coordinator received, would you jump? This shows students quite directly why typical conclusions such as 'if the rope is longer the swing takes longer' are unsatisfactory. As one student puts it: 'my conclusion is of no [expletive] use to him!' (authors' translation). Teacher feedback on the lab report, in our experience, rarely has this effect. Students appreciate that actual filmmaking depends on similar research impacting, e.g., the safety of stuntmen. They conclude that the stunt coordinator needs a report that is convincing (optimally informative, trustworthy and useful) and that theirs is not.

Millar et al. (1994) regard inquiry as the implementation of 'procedural and conceptual knowledge in science' (PACKS). Their PACKS model builds on so-called Concepts of Evidence (CoE), 'certain ideas which underpin the collection, analysis and interpretation of data [that] have to be understood before we can handle scientific evidence effectively' (Gott & Duggan, 1996). The concept at hand is called 'practicality of consequences'. While concepts of reliability and validity are still abstract and remote, our students can consider the costs of implementing their findings, as a step towards developing these targeted concepts. For this, activity 1 uses six design principles:

1. Students carry out their own inquiry. This provides a baseline on students' PACKS.
2. In the first activity they make the usual mistakes so that it can become a constructive 'bad example' – an episode that reinforces how not to address an issue (Kapur, 2008).
3. Students experience the context as realistic and demanding of high quality answers.
4. Students take the roles of 'producers' and 'consumers' of knowledge. The context is suggestive of evaluation criteria such as useful, trustworthy, informative as characteristic of a cogent result.
5. Only basic knowledge and skills are needed, if the inquiry fails it does so in terms of the students' own criteria. They find out for themselves what is needed to do inquiry properly.
6. The activity is 'closed' in that all ought to draw the same conclusions concerning the purposes of inquiry and how to approach them. These conclusions are explicitly formulated as 'rules for doing proper investigations' by the students in their own words at the end of each activity.

4.2.2 Activity 2 – Observation vs. inference with Tricky Tracks

Once students feel a *need* for cogent conclusions, developing a method for constructing and evaluating these is in order. We adapted 'Tricky Tracks' (N. G. Lederman & Abd-El-Khalick, 1998) for this purpose. Young students may regard an observation and its interpretation as one 'fact'. If the possibility of multiple interpretations of a single data set is non-existent, contesting its interpretation makes no sense and inferences need no justification. The claim 'is' the data. Our version of 'Tricky tracks' addresses this by asking students, in turn, to state what they observe in figure 4.1, but without repeating any previous statement. Soon, observations (e.g., '*the shapes are of two different sizes*') are mixed with inferences (e.g., '*the shapes are footprints*'). As all statements are displayed the teacher asks:

- Do you agree with all observations made so far? Why, or why not?
- Can we be sure that birds made these tracks? That they were present at the same time? What makes you think they fought/played/one flew away?
- If you could visit this place what would you do, or pay special attention to? Why?
- What would be a better term than 'observations' for statements we cannot agree upon?

Figure 4.1. Tricky tracks adopted from Lederman and Abd-El-Khalick (1998) to teach the difference between observations and inferences.

Similar situations where a dataset has various acceptable interpretations are explained as common in science. But lacking a unique correct interpretation, we can still seek out and report the best ones available (Lipton, 2003) and draw some tentative conclusions from our data, provided we specify *how we arrive at them* and how *certain we are*:

> *If this is a pattern in loose dirt it is likely that* ***it was recently produced by animals,*** **because this is what footprints look like. Since it consists** *of two shapes that differ in size, it is likely that* ***two animals produced it****. If both animals were present at the same time, we can conclude* **from the usual shape of feet** *that they must* ***have come together in the middle****. There one set of tracks ends. We can firmly conclude that* ***this animal did not leave the scene walking unless footprints were erased****. We may speculate:* ***is it still present, did it fly away, was it eaten or did it climb on the back of the other animal****?*

Generalising this account, a simplified version of Toulmin's (2003) 'model of argumentation' (figure 4.2) provides a method for constructing a cogent conclusion; construct a ***claim*** (e.g., the answer to the research question), moderated by qualifiers and supported by **inferences** (i.e. warrants and backings) based on the *data*. These aspects of arguments have been highlighted similarly, with underlining, italics and bold, in the preceding section.

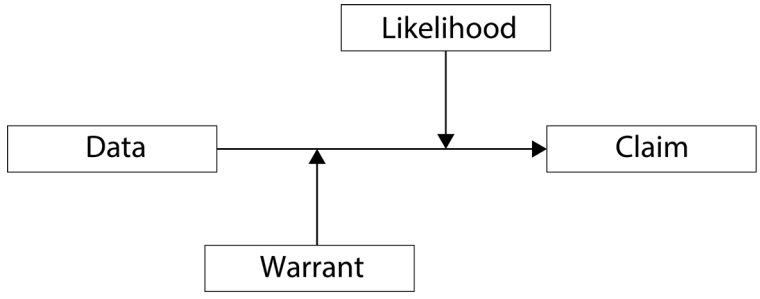


Figure 4.2. A reduced and simplified version of Toulmin's argumentation model is introduced to help structure conclusions.

Students practice the approach by analysing a short online article of the (highly respected) National Dutch Broadcasting Foundation that claims that conclusive evidence has finally been found of the existence of the *Abominable Snowman* or *Yeti*. Students identify the different aspects of the (exceedingly flimsy) argument and evaluate whether they find it convincing.

While students clearly came to distinguish observation from inference implementing this distinction and constructing cogent arguments was no simple matter, requiring further practice throughout the sequence.

4.2.3 Activity 3 – Establishing a relationship in advising the International Swimming League

Inquiry into relationships between variables is especially relevant in school science. In activity 3, students learn that relationships become more convincing if based on (1) more data collected from (2) a larger population, provided that (3) they are obtained through one and the same, appropriate procedure, in which (4) (human) error is avoided. Combining data sets (5) generally enhances trustworthiness, but (6) conclusions apply only to the researched population. The notion that (7) a conclusion is most convincing if it is optimally trustworthy, useful and informative is reinforced.

Reflecting on a 'newspaper article', students consider whether swimmers with relatively long arms have an unfair advantage, warranting the introduction of length classes in swimming. To start investigating the matter and advise the fictitious *International Swimming League* (ISL), students explore the relationship between human body length and arms' width. They measure each other in pairs, then share the data on the interactive whiteboard. A scatter graph gradually appears. They discuss:

- Were the first two data point enough to state a conclusion? Why, or why not?
- How reliable are our data, did everyone measure in the same way?
- What is the relation, if any, between arms' width and body length?
- How certain are we that this relationship really exists? How can we obtain more certainty?
- Is this relationship valid always and everywhere? How can we find out?
- If an additional data set is available should we combine them? What information do we need to decide?

An additional set of over 100 measurements (Figure 4.3) is introduced. The class discusses how it affects the established relationship and previous answers. Next, in the role of ISL Chairperson, students discuss which of the following conclusions, appearing consecutively, is most satisfactory, and why:

1. Taller people have longer arms.
2. There is a relationship between body length and width.
3. Body length and width are directly proportional
4. For people of between 1,50 and 1,90 m in length, conclusion 3 is true.
5. Conclusion 4 is often true, but for one in three people this rule does not apply.

Returning to their researcher roles, students then write a conclusion that is even better than these to the ISL, including also their personal view.

Students responded well, e.g., spontaneously discussing the fit with and meaning of the data pattern as data were still coming in. They identified limitations of the study and proposed appropriate expansions to take into account, e.g., a wider age range and other demographic characteristics.

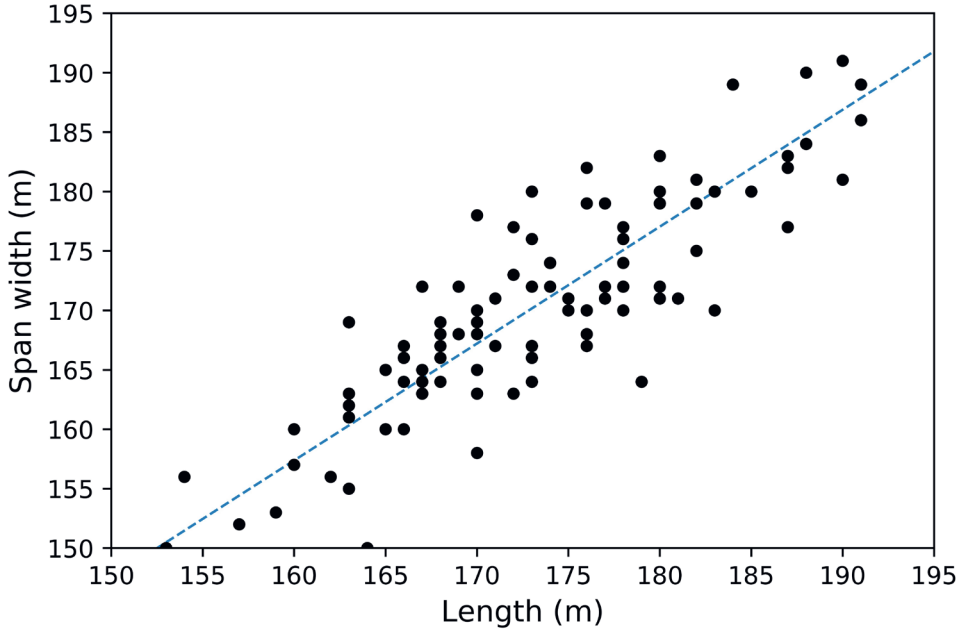


Figure 4.3. A scatter plot of over 100 data points relating human arms' width to body length.

4.2.4 Activity 4 – Data variability and the Fitch Barrier

For students without experience in inquiry repeating a measurement may seem pointless – if you measured correctly, why should it be different? This activity addresses understandings (1)-(7) again, but focuses on repeating measurements and verifying reproducibility. Students learn that variability in the measurements (8) is a natural, unavoidable characteristic that (9) if accounted for makes the conclusion more credible. Students also learn (10) how to deal with outlying data and discuss (11) how many repeats of a measurement suffice.

After his friend's terrible racing accident in 1955, John Fitch invented the Fitch Barrier (Fitch, 1971) consisting of barrels filled with sand. A car crashing into these will decelerate, providing some protection for both the driver and spectators along the road. However if the car slows down too quickly the driver gets hurt – too slowly and the spectators remain unprotected. How many barrels are needed to decelerate the car just right?

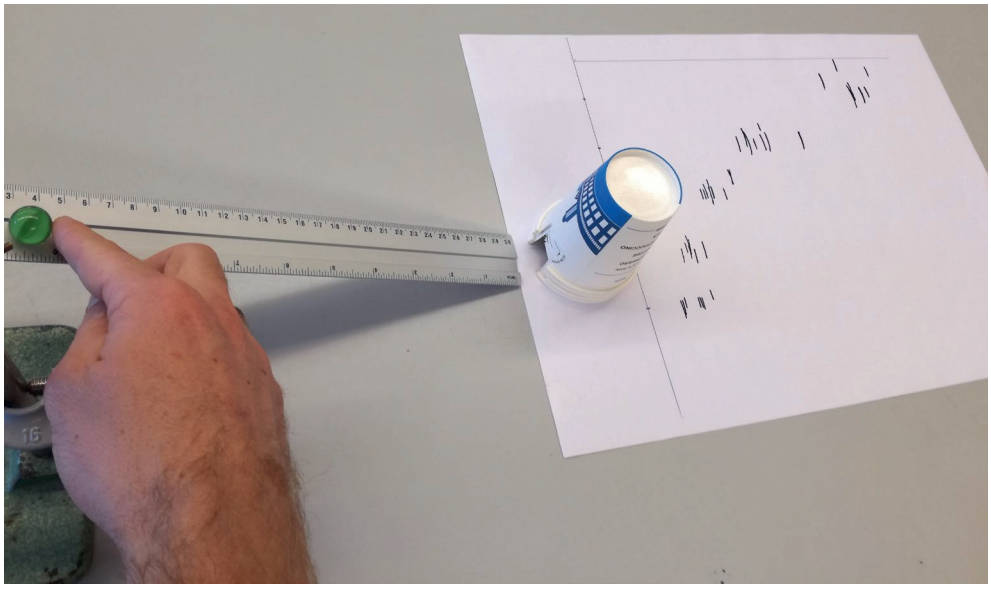


Figure 4.4. In this activity adopted from Farmer (2012), a marble rolls into the stack of cups which then slide across the paper. A line is drawn at that spot. Repeating the procedure reveals variability in the measurements.

Students investigate this in an activity adapted from Farmer (2012). A marble rolling down an incline into stacked cups (with a hole in the side) will slide some distance and stop. The cups model the barrels, the marble represents a car. Possible extensions of the inquiry are easy to envisage. Students determine the relationship between the number of cups and their sliding distance. During the experiment the teacher monitors and asks questions like:

- Do you get the same results in repeated measurements? Why do the outcomes keep on changing? Does that mean you are not measuring properly? Can you reduce the spread?
- For a given number of cups, your results are roughly but not precisely the same. What is a trustworthy way to report this? How do you report the (real) value of the sliding distance?

Students became aware that no matter how well they tried to repeat the measurement, the stopping distances always varies, even though both it and its absolute variability become smaller with more cups (Figure 4.4). Concept cartoons (Figure 4.5) were discussed to decide how to deal with outlying data, and how many repeats of a measurement are needed to ensure a reliable value. Students also evaluated their inquiry by discussing issues of reporting reproducibility and data collection: they compared the results among the teams and reported similarities and differences. They discussed how trustworthiness increases if all data are reported as well as the method of collection. The teacher assisted in establishing the nonlinear relation by suggesting to students to look at changes when

doubling the independent variable instead of using equal increments, something students are unlikely to figure out unaided.

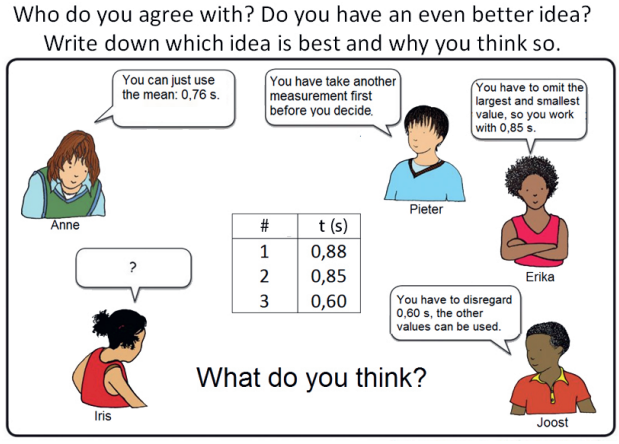


Figure 4.5. An adapted concept cartoon (Keogh & Naylor, 1999) is used to discuss the difficult concept of dealing with outliers.

Students noted without help that reporting the average measured values would not suffice here. Although the averages contribute to establishing the relation between mass and stopping distance, they argued that guaranteeing the safety of the people requires that the extreme measurements are also reported. Activity 4 uses an additional design principle:

7 Developing Toulmin's model and understandings (1)-(11) is an explicit aim of learning in Activities 2-4. Understandings (1)-(11) invoke a range of the CoE specified in Gott et al. (2003) including, e.g., concepts such as 'datum', 'measurement', 'variability', '(in)dependent'. 'variable', 'repeat', 'reproduce', 'average', 'fair test', and 'line of best fit'.

4.2.5 Activity 5 – Practicing what was learned and NASA's Escape Pod

In Activity 5 students practice what was learned and consolidate their learning by reflection in helping NASA design a new escape pod for astronauts. The computer model designing the pod requires very accurate input, especially on factors influencing the frictional force. The pod is modelled as a paper cone (e.g.: Mooldijk and Savelsbergh (2000), measurements involve its falling (figure 4.6). Potential factors are identified and allocated to research teams for further study: distance fallen, mass, diameter, top angle of the cone. During the investigation the teacher asks supporting questions about the cogency of students approach and the use of CoE's such as 'fair testing'.

Since only adequate work is to be included in the final report, the teams evaluate each other's contributions to judge whether inclusion is warranted. Each team uses its own checklist of evaluation criteria drawn from the '*rules for doing proper investigations*' written up in the preceding activities. Thus was explored whether the students apply the appropriate CoE adequately in their own work and can recognize this in the work of others.

The quality of students' work varied from two groups designing an Arduino-based electronic timing device eliminating response time, to a group who forgot to measure the cone's diameter in exploring the influence of its frontal area on falling time. Despite the variety in approaches, the vast majority of data collection procedures improved considerably to previous student practice. Students accounted for their choices in terms of the reliability of the data, showing understanding of the relation between research procedure and quality. The ability to analyse data and draw the most informative conclusion remained limited. This final activity uses the following design principle:

8 Activity 5 is designed for students to consolidate previous learning as they engage in inquiry. They summarize and apply insights on how to do inquiry properly, and reflect on how they developed these insights. Students and teacher learn whether the intended understandings have been fully developed or require further clarification.

Figure 4.6. Students drop paper cones with different frontal areas and measure the falling time using both a stopwatch and their mobile phone camera.

4.3 Conclusion

We wanted students to see why their usual conclusions in inquiry are unsatisfactory by scientific standards. Since students do not yet have these standards, they were asked to consider if their conclusion was good enough if their personal safety depended on it. They realised that it was not, as it was not optimally informative, trustworthy and useful. Student then learned that a conclusion is in fact one interpretation of the research data while many tend to be possible. A conclusion in inquiry therefore should be an argument, consisting of a claim, the data, and the statements that link the two, providing support for the given claim.

Inquiry in science is directed at finding the best possible claim given the circumstances, i.e. the most informative, trustworthy and useful conclusion. In order to convince themselves and each other that a conclusion is the best available, scientists use a range of understandings. Eleven of these, all about optimizing the quality of the data and their interpretation, were developed in activities 3 and 4. Throughout the sequence, students drew up 'rules for doing proper inquiry'. In the final inquiry they used these to design and report an investigation of factors influencing air resistance and to evaluate the reports of others.

As expected, students did not develop straight away a high proficiency in applying Toulmin's model of argumentation or in applying the eleven understandings of evidence. They did, however, come to apply more appropriate data collection procedures, choose a wide range of many values for the independent variable, repeat measurements and calculate averages, take into account and report data variability, deliberately try to reduce or eliminate error, and consider various interpretations of any data set. Importantly, they clearly understood why they should do all of this. This, in our view, provides a useful starting point for more challenging kinds of scientific inquiry.

4.4 Reflection and next step(s)

With a detailed description of the TLS, we address in the next chapter areas of concern 1 and 2. In the first activity of the TLS we try to engage students in inquiry that they see as relevant and worthwhile. We try to attach personal relevance to the task so that they become interested in finding a useful and trustworthy answer and are stimulated to take responsibility for finding these.

5. "*Would you dare to jump?*" Fostering a scientific approach to secondary physics inquiry

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

Even secondary school students who know the rules and procedures for doing proper scientific inquiry often use these only when prompted, as if they fail to see the *point* of doing so. This qualitative, small-scale developmental design study explores conditions to address this perennial problem in school science inquiry. Dutch students (N=22, aged 14-15) repeatedly consider the quality of their work: (1) in a conventional, guided inquiry approach, (2) by evaluating their conclusion in terms of the contextual purpose of the investigation, (3) as consumers of knowledge facing the (hypothetical) risk of applying the findings in the real world. By gauging students' level of confidence in the trustworthiness of their results, we established that, while each confrontation instigated some students to (re)consider the quality of their inquiry, the final stage had the greatest impact. Students came to see that finding useful and trustworthy results is essential and more likely if scientific standards are applied. Using a validated rubric the scientific quality of their inquiries was described, weaknesses identified and compared with the improvements students themselves proposed for their inquiries. While the students' proposals were expressed in non-specific terms these align with a scientific perspective. Students now *wanted* to find useful and trustworthy answers by exploiting the power of scientific standards. In enabling students to engage successfully in basic scientific inquiry, finding ways to establish students' mental readiness for attending to the quality of their scientific claims, and of personalised scientific criteria for their assessment, is indispensable.

5.1 Introduction

Practical work refers to activities in which students manipulate instruments and materials to answer a research question (Millar et al., 1999). It is frequently used to achieve two broad aims in science education: 1) to help students develop a proper understanding of the relation between scientific theory and practice, and 2) to become competent in conducting their own scientific research (Abrahams, 2005; Hodson, 2014; Hofstein, 2017; Hofstein & Kind, 2012; Millar, 2004; Millar et al., 1999). This paper focuses on the second of these aims. Despite many decades of research and development, practical work in most of today's classrooms still involves students doing no more than following up on detailed instructions (Abrahams & Millar, 2008; Holmes & Wieman, 2016, 2018; Wieman, 2015). When instructed to do so, the students repeat measurements sufficiently often, calculate averages correctly, apply appropriate instruments, use suitable tables and graphs, etcetera. But as soon as we stop telling them what to do, they stop doing so, and are unable to find valid and reliable answers by themselves (Millar, 2004). In other words, we have been unable to use practical work effectively to enable students to engage in basic scientific inquiry independently (Abrahams, 2011; Abrahams & Millar, 2008; Hofstein, 2017; Hofstein & Kind, 2012; Hofstein & Lunetta, 2004; Lunetta et al., 2007).

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

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

This paper is based on the assumption that before we can expect students to make desirable choices in inquiry, we will have to teach them the value of that *scientific quality.*

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

5.2 Theoretical framework

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

5.2.1 A model of the knowledge applied in practical work – PACKS

Practical work should be a minds-on activity characterised by students' use of their *Procedural and Conceptual Knowledge in Science (PACKS)* (Millar et al., 1994). Figure 5.1 presents the PACKS model and the different types of knowledge (**A**-**D**) that influence researchers' decisions in the different stages of an inquiry (Millar et al., 1994). The model distinguishes knowledge of (**A**) the nature and purpose of the inquiry, (**B**) relevant content, (**C**) required manipulative skills and (**D**) evaluating scientific evidence. Consideration of the quality of the research involves, in the first place, application of type **D** knowledge, comprising of awareness and use of criteria involved in the construction and evaluation of scientific evidence. These criteria include, *i.a.*, an operationalization of the *Concepts of Evidence (CoE)*. These are concepts such as fair test, experimenter bias, range, median, precision and measurement uncertainty that underpin the more abstract concepts of reliability and validity (Gott & Duggan, 1996; Gott et al., 2003). Using a scientific approach in practical work entails the conscious and adequate use of this type of knowledge in finding and evaluating answers to the question: *At this point, what needs to be done to achieve the*

best possible result in this investigation in the given circumstances? 'Best possible result', that is, in terms of the scientific goal of describing, explaining and predicting events and phenomena as precisely and accurately as possible.

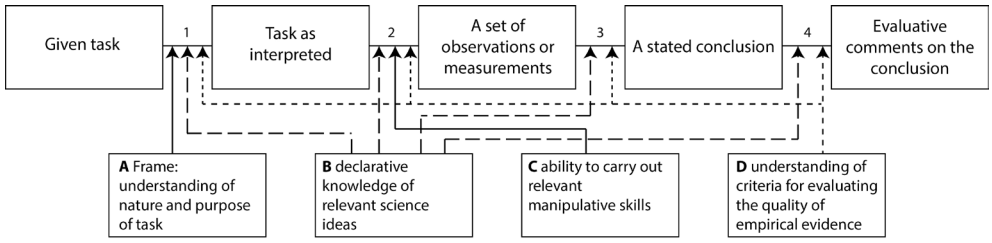


Figure 5.1: The Procedural and Conceptual Knowledge in Science Model (Millar et al.,1994) illustrates how different types of knowledge, on the left, influence decisions made at various stages of an inquiry.

5.2.2 Understandings of evidence

Rather than evaluating the presence of isolated concepts, we proposed to consider collections of loosely interrelated CoE that constitute overlapping '*Understandings of Evidence*' (UoE) (Pols, Dekkers, & de Vries, 2022a). UoE express properties of the evidential information at a particular stage, or procedures for constructing that information, as well as prescriptions for enhancing or assessing informational quality. The UoE delineate knowledge a researcher has and applies in constructing an optimally reliable and valid inquiry. They are the common understandings by which researchers evaluate and judge the quality of their own research and that of others, the norms and standards they use to determine how well the empirical data support the researcher's claims. An Assessment Rubric for Physics Inquiry (ARPI) was constructed and validated by Pols et al. (2022a) that allows for assessment of a student's UoE based on his or her observable inquiry actions and research report. The instrument distinguishes 19 UoE distributed across six phases of inquiry: (1) Asking questions, (2) Design, (3) Methods & procedures, (4) Analysis, (5) Conclusion and evaluation, (6) Peer review. For each UoE, indicators for the lowest, intermediate and highest levels on a five point scale are provided, see Table 5.1, where levels in between are assigned when a student outperforms the lower level but not fully attains the higher level. Depending on the openness of the inquiry, the precise task and the specific learning goals, specific (clusters of) UoE can be selected for assessment purposes. For instance, in structured inquiry (Table 5.2), only ARPI's clusters (4)-(6) can be assessed as students are using a given research question and method.

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

Table 5.1: Illustrative excerpts of the Assessment Rubric for Physics Inquiry (Pols et al., 2022a). Five levels of competence are distinguished for each Understanding of Evidence described on the left. Descriptors for lowest, intermediate and highest levels are specified, where intermediate levels can be assigned when a student outperforms the lower level though not yet fully attains the higher level.

UoE		**Level of competence**		
Phase	**The researcher understands that:**	**4**	**2**	**0**
Method & procedure	8: measured values will show inherent variation and the reliability of data must be optimised, requiring repeated measurements.	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.
Analysis	12: Data require appropriate methods for analysing and describing them.	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.	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.

5.2.3 Guided Inquiry

While developing understandings of scientific inquiry requires that students (be given the opportunity to) take agency and learn from the consequences (Hodson, 1992), inexperienced students conversely need support and structure. In terms of student input and choice, *guided inquiry* offers a balance (Banchi & Bell, 2008; Tamir, 1991) that is appropriate in this study. Table 5.2 shows that the research question is posed by the teacher but the answer is unknown by the students beforehand and they decide on the procedure. Students will all attempt to answer the same research question. However, depending on their ideas about evidence and their understanding of what constitutes 'good science' (Gott & Duggan, 1996), they will make different decisions. Students will thus differ in how they

answer the research question, and in the scientific quality of that answer. Since all students can consider the quality of their own and each other's work in terms of the same scientific purpose, this quality can become the focus of attention rather than the details of subject matter, the experimental setup or the data analysis.

Table 5.2: Tamir (1991) distinguishes four levels of inquiry, depending on the information provided to the student. Guided inquiry balances the teacher's support with students' independency to organize the research as they see fit.

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

5.2.4 Context-based approach

A context-based approach is advocated in various curricula including the Dutch physics curriculum (Bennett, Lubben, & Hogarth, 2007; de Putter-Smits, 2012; Netherlands Institute for Curriculum Development, 2016) as intrinsically more authentic, stimulating and interesting. Students are assumed to put more effort into learning content that is perceived as relevant because of its context (Kortland, 2007). However, if a context merely serves to teach difficult concepts, students quickly lose interest (Kortland, 2007; P. Lijnse, 2014, p. 157) and forget the context (Molyneux-Hodgson, Sutherland, & Butterfield, 1999).

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

5.3 Method

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

5.3.1 Research design

Informed by the literature on teaching inquiry in science education, practical work and context-based approaches, we developed an intervention consisting of three stages. Each

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

5.3.2 Participants and educational context

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

The teacher, also the first author of this paper, had 9 years of teaching experience in physics at secondary school. Well aware of the challenges involved he had conducted several in-service and conference workshops on practical work and teaching scientific inquiry (Pols, 2021b). As advocated in the literature, teachers' research of their own practice is an authentic way to study 'what goes in the school laboratory' (Hodson, 1990; Hofstein, 2017) and the students' behaviour and constructed perceptions and understandings (Hofstein and Kind (2012). It has the potential to close the research-practice gap (Bakx et al., 2016). *In every thesis an Easter egg should be hidden, just like this one. Just as an acknowledgment of the author to the reader, showing the appreciation for carefully reading the whole manuscript.*

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

some (implicit) understandings of inquiry these result from closed, 'cookbook' experiments that tell students precisely what to do.

5.3.3 Educational Design

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

The teacher's role during the first part of the intervention was modest. He was to explain the task, emphasizing that the stunt was to be filmed in a single take. Students were then expected to devise the experiment as they see fit. If students had questions related to the given task, to the physics involved or to the use of (more advanced) research methods and instruments (knowledge types **A**-**C**), these were to be answered directly so as to reduce the chance of cognitive overload. This would allow students to focus on knowledge type **D** only (Johnstone & Wham, 1982; van den Berg, 2013). If students had questions addressing knowledge type **D**, or if the teacher observed errors in, e.g., controlling variables, these issues were to be discussed on the spot.

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

Apart from the use of a film clip, this approach so far is conventional. Since inexperienced students tend to be brief and superficial in their construction, justification and evaluation of conclusions in inquiry, and the intervention so far does not affect this, no serious consideration of the quality of the answer to the research question was expected.

A conventional practical would end here, with a brief lab report on what factors affect the period of the swing (and possibly a teacher explanation of the appropriate formula).

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

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

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

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

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

2 Once students consider the quality of their answer to the research question, what aspects of this perceived quality align with scientific quality, and which aspects are missing?

5.3.4 Instruments and data collection

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

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

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

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

(iv) Reflection form During the third stage, the second lesson, after the whole-class discussion on 'would you dare to jump', each student team answered the following questions in writing:

1 What would you like to change in your investigation? Why?
2 What do you want to achieve with that change?
3 What makes an investigation and the written report trustworthy?

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

(v) Audio recordings of the second lesson Again, the teacher recorded classroom talk during the entire second lesson with an audio voice recorder. Where most of stage 1 consisted of work in small teams, this stage included a whole-class discussion that introduced the change of role from producer to consumer of knowledge. It also included conversations during whole-class and small-team reflective activities to evaluate the quality

of the inquiry. The data provide information about the effects of the change of role, the students' self-evaluations and their ideas for improvement.

5.3.5 Data analysis

ARPI was used to describe, analyse and rate the students' approach in the first stage, based on the choices they made in designing and executing their inquiry. Relevant information for each targeted UoE was gathered from the SGO, the letter to the stunt coordinator and the audio recordings of the first lesson. Analysis of the three data sources revealed what choices students made (e.g., regarding the number of repeated measurements) and whether they consciously substantiated these choices (e.g., with a statement such as "*since the spread in measurements is small, three repeats suffice*"). The ARPI descriptors were used to assign attainment levels to the teams and score the quality of students' actions and substantiations. For instance, for UoE 8 (Table 5.1) we first analysed whether students collected a single measurement (level 0), or took repeated measurements (level 2). We then investigated whether students provided a substantiation of that decision (level 4). Levels 1 or 3 were assigned when students outperformed the lower level, but did not fully reach the next level, e.g., level 3 could be assigned when measurements were repeated but an incomplete or mediocre substantiation was provided.

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

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

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

To study whether the switch in perspective changed the students' perception of the usefulness and trustworthiness of their inquiry, we analysed first the level of confidence students had in their results, as expressed in the letter to the stunt coordinator. We allocated *low (-)*, *intermediate (0)* and *high (+)* levels (or *(?)* if not expressed). Subsequently we analysed the audio recordings of the stage 3 with a focus on students' reactions and arguments when asked whether they would dare to jump. We compared their views in the letter with these verbal reactions and arguments.

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

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

5.4 Results

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

Table 5.3: Number of teams (N=11) per competence level for each UoE on a 5-point scale from lowest (0) to highest (4), on average in SGO and letter. Class average level in grey. Number of teams whose UoE could not be determined in final column.

UoE			Level of competence					
Phase	**no**	**The researcher understands that**	**0**	**1**	**2**	**3**	**4**	**No score**
Research Question	2	The inquiry is an attempt to establish the relationship (or lack of one) between an independent variable and a dependent variable.	0	0	5	1	5	0
Design	4	The research question should be answerable with the devised experiment.	6	0	2	2	0	1
	5	Other variables can affect the dependent one, therefore a fair test is needed, keeping these variables constant.	2	1	5	0	3	0
	6	It is important to choose suitable instruments and procedures to get valid data with the required accuracy and precision.	10	1	0	0	0	0
	7	(Human) Errors and uncertainties may occur and precautions are needed to minimize or avoid them, ensuring reliability.	3	6	2	0	0	0

Method & Procedure	8	Measured values will show inherent variation and the reliability of data must be optimised, requiring repeated measurements.	2	1	8	0	0	0
	9	The range of values of the independent variable must be wide enough and the interval small enough to ensure that a potential pattern is detectable.	3	0	3	2	1	2
Analysis	12	Data require appropriate methods for analysing and describing them.	0	2	3	2	3	1
	13	An optimally informative answer to the research question requires a description of relationships in as much detail as possible. Quantitative descriptions are more detailed than qualitative ones.	3	1	3	2	0	2
Conclusions & Evaluation	14	A complete, clear, substantiated and useful answer to the research question must be formulated.	3	3	0	4	0	1
	15	The reliability of the dataset is to be accounted for by considering how well each datum was measured and the reliability of the established relationship.	1	3	5	0	0	2
	16	The validity of conclusions does not go beyond the data available. Therefore limitations to the validity of the claim should be expressed.	2	4	4	1	0	0

5.4.1 Students' inquiry from a scientific perspective

For each of the twelve UoE of ARPI, attainment levels of each student team were assigned on the basis of their SGO inquiry reports and letters to the stunt coordinator. The results are shown in Table 5.3. The student teams' operationalisation of their inquiry is analysed as follows:

ARPI Phase: Posing questions

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

ARPI Phase: Design

UoE 4: A relation between the experiment and the research question was often not specified. While most teams chose generally suitable instruments and procedures for measuring relevant quantities, a systematic, structured approach tended to be absent. The most extensive description was given by team G9: 'in order to see how mass influences the swing time, seven different weights (20-100 g) were used'. Another more extensive description, in the letter of G1, is presented in Figure 5.3.

UoE 5*:* Teams mostly identified variables that could potentially influence the 'swing time' and understood that therefore, these needed to be controlled (i.e., kept constant). Several

failed to adequately operationalize this understanding, *e.g.*, various teams increased the weight of the pendulum by hanging additional weights below one another. The ensuing discussion with the teacher showed that they understood 'fair testing' (change only one variable at a time to establish its effect) but failed to notice that their way of increasing the weight also increased the pendulum's length.

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

ARPI Phase: Method & procedure

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

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

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

ARPI Phase: Analysis

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

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

ARPI Phase: Conclusions & Evaluation

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

G3: The lighter the weight, the shorter the swing time, so it seems.

G6: The difference per rope (material) is minimal, but of importance for timing the perfect jump.

G10: The bigger the (starting) angle, the longer the swing time, but noticeably only from 40^{o} onwards. It doesn't differ much, but it is clear.

These qualitative conclusions did not meet scientific requirements, they were insufficiently informative and not useful from that perspective in the given context. Especially in cases where the relation is not present or measurable (rope material, mass) or not evident (angle), students had difficulties in describing the effect of the variable under investigation.

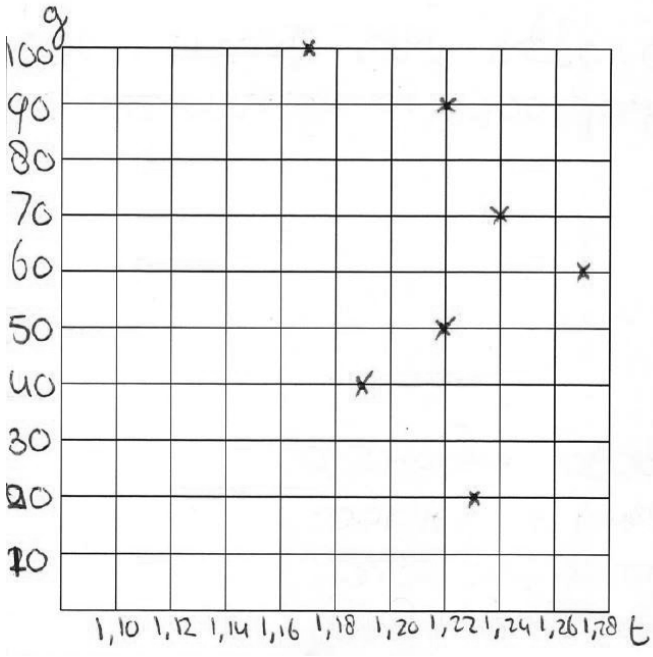


Figure 5.2: Students interpreted their well-presented data as showing that an increase in mass results in a larger period, although the variation in the measured period is within the margin of error and this inference unwarranted.

5.4.2 Stage 1: Student's initial perspective of their inquiry

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

Exchange with G9:

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

Exchange with G5:

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

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

In line with the findings in Table 5.3, most carried out the inquiry as is often reported in the literature: quickly and without explicit consideration of the quality of data. While some students genuinely believed they tried as hard as they could, they did not feel the urge to ask the teacher for better instruments or methods to determine the swing time.

5.4.3 Stage 2: Students' perspective of inquiry in the letter to the coordinator

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

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

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

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

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

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

Figure 5.3: Two exemplary letters to the stunt coordinator in which students explain what has been done and found and whether they have confident in the quality of their own inquiries.

Four of the eleven teams (G1,G2,G4,G9) stated that they were not confident that the inquiry findings could be used, four teams had some confidence (G3,G7,G10,G11) and two teams (G5,G8) explicitly stated they did have confidence in their own findings. Team G6 did not mention its level of confidence. Notably, their lack of confidence was not due to a lack of effort. The teams did feel they tried to produce quality research (as exemplified in the underlined sections below) but that they encountered 'insuperable', externally attributed problems:

G6: We measured 4x per rope to increase the accuracy of the measurement. It is hard to measure accurately, but we tried the best we could. (…) We hope to have helped you.

G9: We tried our utmost, but because we did not have equipment to measure very precisely, we are, unfortunately, not confident that our research findings are useful.

Note that the two letters presented in Figure 5.3 also present a (justifiable) lack of confidence.

5.4.4 Stage 3: Students' view of inquiry in the reflective discourse

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

Teacher:	Suppose you are the stuntman standing on the edge of the ship, you have a 12 meters long, 5 centimetres thick rope in your hands and you have to jump soon. Just before the stunt, you have read how the stunt should be performed. The stunt is based on your own reports and investigations… Would you dare to jump?
Lisa (G9):	(interrupts) No.
Teacher:	Why not?
Lisa (G9):	It is not measured with proper equipment, it is based on us… you do not know how and what, exactly.
Teacher:	You think your measurements are not adequate enough?
Lisa (G9):	No.
Teacher:	And that is due to the equipment?
Lisa (G9):	And ourselves, you cannot start from the exact same starting point each time. And the equipment is not good, better equipment is required.
Teacher:	So what do you suggest? What do you want to improve?
Lisa (G9):	Every time using the same point for your measurement.
Teacher:	The angle at which you start you mean?
Lisa (G9):	Yes, and where you start and stop timing.

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

Teacher:	Who would dare to jump?
Thim (G10):	Sure, why not? (other students are laughing)

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

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

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

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

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

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

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

5.4.5 Students' written reflection on what is learned

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

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

> G1: Use a larger longer rope, this increases the swing time and makes it therefore easier to accurately measure the time. Use a sensor to measure when the swing is released and stops when the swing is at the other side. This way you don't have to deal with reaction time and thus results in a more accurate measurement. Attach the triangle ruler to the setup in order to measure angles accurately.
>
> G3: We would like to use professional equipment for obtaining measurements. We probably did not measure and calculate everything perfectly resulting in findings that are not quite right. What we want to achieve with this is that we can optimize our conclusion and the stunt can be performed in a safe way.

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

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

instruments (UoE6).

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

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

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

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

5.5 Discussion

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

5.5.1 Answers to the research question

Stage 1 of the practical involved the deceptively simple physical pendulum (Matthews, 2001) but deviated from the conventional 'cookbook' exercise to confirm the formula relating length to period. As suggested by various scholars, we gave students more agency of their inquiries (Crawford, 2014, p. 527; Hofstein & Kind, 2012; Zion & Mendelovici, 2012). We encouraged them to forward their own ideas about factors that might influence the period, and to study these as they saw fit. Reducing the cognitive load in terms of knowledge of types **A**, **B** and **C** of the PACKS model allowed students to focus on the aspects involved in knowledge type **D**: their use of criteria involved in the construction and evaluation of scientific evidence. We observed, however, that context-based, guided inquiry and explicit self-evaluation of the research quality did not sufficiently affect the students' intent (RQ1a). For example, when asked to consider the quality of their answer to the research question (Q: '*Would the stuntman have confidence in the research?*'), students understood that quality to be adequate in a scientific sense (A: '*Yes, because it is scientific*'). The students' words and actions did not sufficiently reflect the understandings that are required to render scientific adequacy to evidence in inquiry (Table 3). For example, they chose the first inquiry methods and approaches that came to mind without searching or asking for better alternatives. 'Better', that is, in terms of criteria they themselves formulated later on in stage 3, but not during stage 1. With very few exceptions, the guided and contextualised character of the activity does not sufficiently foster students' awareness of the value of a scientific approach (RQ 1a), confirming findings of e.g. Molyneux-Hodgson et al. (1999).

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

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

collectively, to identify these ordering criteria and to interpret them as making the answer more useful and trustworthy, therefore preferable.

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

5.5.2 Implications

According to the literature, students in inquiry (seem to) act almost without thinking, (seemingly) indifferent to establishing a valid and reliable answer to the research question, or ignorant of how to obtain it. This study shows, however, that even if students appear interested, motivated and engaged: they fail to see the point of obtaining better answers and lack criteria for evaluation of the quality of such answers. In making practical work more effective and enabling students to engage in basic scientific inquiry (Abrahams, Reiss, & Sharpe, 2013; Hodson, 2014; Hofstein & Kind, 2012) we direct students' attention to the value and purpose of scientific investigations. The question of why some answers are better than others, and what is meant by 'better' in science, appears to be a useful starting point for learning the methods and techniques scientists apply to optimize the quality of their inquiries. As shown, appealing to students' empathy and encouraging them to develop personally relevant criteria is one way to do so. The combination of context, reflection and a change of perspective from producer to consumer of knowledge contributes to an educational design that accomplishes this. While the intuitive concepts 'trustworthiness' and 'usefulness' are not necessarily fully developed in a scientific sense, they align with and can be developed further into the more fundamental but abstract concepts of reliability and validity.

This study has implications for integrating argumentation into inquiry, advocated by influential authors (Erduran & Jiménez-Aleixandre, 2008; Gott & Duggan, 2007; Newton et al., 1999; Osborne, 2013) but scarce in terms of empirical studies attempting it (Driver et al., 2000; Erduran & Jiménez-Aleixandre, 2008; Watson, Swain, & McRobbie, 2004). We have argued elsewhere that conducting inquiry can be interpreted as the construction of an optimally cogent argument in support of an optimally informative claim on the basis of

optimally valid and reliable data (Pols et al., 2022a). Engaging in argumentation requires students to have a notion of what counts as scientifically cogent, i.e., of what makes some answers to research questions better, in a scientific sense, than others. This study provides an example of a starting point for developing these notions and satisfying the preconditions for students engaging in argumentation. We have provided an example of students' successful argumentation in establishing the most informative answer to their research question.

5.5.3 Limitations and future research

More research is needed to explore how the learning effects in this intervention can be consolidated and utilised in the further developments described above. As a first step, the collection of UoE in Table 6.3 has been validated as a set of norms and standards by which the quality of virtually all students' inquiry in physics can be assessed. This set of UoE is suitable in guiding student-researchers in developing or evaluating that quality, and in argumentation aimed at the construction or evaluation of the scientific cogency of a researcher's claims. Developed and validated with physics students at BSc level, the next step will be to develop learning pathways for levels between that of the current study and university level. As a starting point, a teaching-learning sequence was developed targeting a range of the UoE that integrates the current intervention. It explores the further development of inexperienced students' intuitive concepts in inquiry learning and argumentation.

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

5.6 Conclusion

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

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

5.7 Reflection and next step(s)

Now that we seemingly have created a drive in students to produce a quality answer to the research question we can exploit it to teach them what such an answer entails and how it can be produced. In doing so we mainly address areas of concern 3 and 4. We investigate more deeply how our formulated design principles contribute to learning and fostering students' critical attitude. Moreover, we investigate the effectiveness of the TLS in terms of learning outcomes, and expand our premature pedagogical theory of teaching scientific inquiry (elaborated on in chapter 3).

6. Integrating argumentation in physics inquiry: a design and evaluation study

This chapter is submitted in adapted form.

This small scale, qualitative study uses educational design research to explore how a focus on argumentation may enable students to engage in inquiry independently. We surmised that if students understand that inquiry can be regarded as the construction of a scientifically cogent argument in support of a claim, they may develop their own reasons for adhering to scientific criteria. 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 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, primarily, classroom observations (N=23, aged 14-15), students' answers to worksheets and self-reflection questions we explored whether the design principles resulted in intended students' actions and attitudes. We studied whether the activities indeed 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. In doing so, they gradually uncovered key characteristics of scientific evidence, understandings of what counts as convincing in science, and why. Students did not yet develop the traditional inquiry skills in these activities, but developed a cognitive need and readiness for learning these. 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. The study contributes to our understanding of how to enable students to successfully engage in inquiry by extending the theoretical framework for argumentation in teaching inquiry and developing a tested educational approach derived from it.

6.1 Introduction

Enabling students to engage in independent scientific inquiry is a highly valued but seemingly elusive goal of science education (Abrahams & Reiss, 2012; Hodson, 2014; Hofstein & Kind, 2012; Hofstein & Lunetta, 2004; Holmes & Wieman, 2018; Kozminski et al., 2014; Lunetta et al., 2007; Millar & Osborne, 1998; Next Generation Science Standards, 2013). To attain this goal in secondary school physics education, students often engage in quantitative physics inquiry (QPI) – the type of inquiry in which a quantitative relation between variables is investigated. In small teams of 2-4, students manipulate instruments and materials to answer the given research question (Millar et al., 1999), which in QPI often is of the form "What is the mathematical relationship between X and Y?" Yet, even after many decades of research and development, we hardly seem to have made progress in attaining this goal. Students do not use the rules and procedures for obtaining optimally reliable and valid data in inquiry unless they are explicitly instructed what to do (Kanari & Millar, 2004; Millar et al., 1999; Pols et al., 2021). Even when they seem to be motivated, interested and able, students rarely independently select an adequate data range, number of repeated measurements, optimally suitable measuring instruments, or make other methodological decisions adequately (Hodson, 1990; Lubben & Millar, 1996; Millar et al., 1999; Tasker & Freyberg, 1985). They hardly seem to think about what they are doing and why they do it in that particular way (Holmes & Wieman, 2016, 2018). Nor do they consider how they could improve the quality of the outcomes (Abrahams & Millar, 2008; Holmes & Wieman, 2018; Pols, 2020a; van den Berg, 2013). Ergo, without guidance, students seem unable to plan and conduct an experiment adequately, to collect reliable data and to produce an informative conclusion.

In this study, we propose an explanation and test its implications with an educational design derived from it. We assume that if students know what to do and know how to do it well enough, but still do not do so, they probably fail to understand the point of doing so. Students may know that the purpose of research is to answer the research question, but may not understand or appreciate yet that only the *best possible* answer is good enough – that is, the most informative, reliable and valid answer to the research question within the given constraints and limits imposed by the feasibility of obtaining it (Pols et al., 2022a). As Osborne (2014b) noted, '*it is not just a matter of knowing how to get reliable data, but also why reliability and validity are important*'. It is therefore argued that progress in enabling students to devise and conduct a high-quality QPI is likely to be made if students understand that the purpose of inquiry is to produce a scientifically cogent answer to the research question, if they have sufficient reason to obtain it, and if they can use argumentation adequately to guide their inquiry towards this kind of answer (Millar et al., 1994; Pols et al., 2021, 2022a; Pols, Dekkers, & de Vries, 2022b).

In our previous study (Pols et al., 2022b), we described one way to introduce the notion that in inquiry only the best available answer is satisfactory and studied how this notion affected students' critical attitude. Even though the QPI was situated in a context that ought to demand a high-quality answer, students (aged 14-15) conducted the inquiry as described above: seemingly without considering the accuracy and adequacy of the obtained results. But once the students were made to consider the quality of their QPI – when asked whether they would be willing to be subjected personally to application of their research findings – their views changed. Even by their own standards, they no longer accepted the quality of their results. Their understanding of the purpose of the inquiry seemed to have changed from '*finding an answer to the research question – any answer will do*' to '*finding the most trustworthy and useful answer obtainable with the means and the time available to us*'. However, the knowledge required to figure out how to produce a more meaningful answer to the research question still needed further development. Moreover, it remained unclear whether students' mental readiness for learning how to do inquiry and their critical stance in this regard would persevere.

In inquiry, students need motivation and drive to invest enough time and energy to obtain an answer of sufficient scientific quality. To obtain motivation and drive they need to understand why that quality is important, what characterizes it, and how it can be obtained. These are the very understandings that guide scientists in their own research and the evaluation of that of others. They are the guidelines by which scientists construct and judge the *cogency* of a claim in view of the obtained scientific evidence. Our approach is to, first, have students consider that the quality of their data, forming the basis of scientific evidence, is crucially relevant. Then, to focus their attention on several of the common understandings that are used to gauge that quality – understandings that scientific arguments in support of research claims are based upon. Finally to 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, to explore what understandings students develop in those activities, and to determine the contribution of those understandings to students' ability to engage independently in QPI.

6.2 Theoretical framework

Following notable experts (Erduran, 2018; Erduran et al., 2004; Gott & Duggan, 2007), we discuss below the central role of argumentation in scientific inquiry, specifically in physics. We describe the structure of an argument using the Toulmin model (Toulmin, 2003), and its content using the *Procedural and Conceptual Knowledge in Science* (PACKS) model (Millar et al., 1994). *Understandings of Evidence* (UoE) (Pols et al., 2022a), the insights a researcher

uses to produce a cogent argument in support of a claim, are presented as the targeted learning goals.

6.2.1 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*' (Kelly, 2014, p. 329). Argumentation plays a central and decisive role in the scientific enterprise and therefore deserves an equally important role in science education (Erduran & Jiménez-Aleixandre, 2008; Gott & Duggan, 2007; Newton et al., 1999; Osborne, 2013). The central role of argumentation in science is especially evident in scientific inquiry (Kelly, 2014). Even though the researcher may not yet know what peer criticism will be received, much thought and effort is invested in making the study's claim 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 is done by, among others, 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 (American Psychological Association, 1983; Chalmers, 2013; Oreskes, 2018). In this process, 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 (Gott & Duggan, 2007; Toulmin, 2003; Woolgar & Latour, 1986). 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 (Gott & Duggan, 2007; Pols et al., 2022a).

6.2.2 The structure and content of scientific argument

In Toulmin's model of argumentation (Toulmin, 2003) an argument consists of field-independent as well as field-dependent elements. Gott and Duggan (Gott & Duggan, 2007) adapted the model to the 'field' of secondary science inquiry, see figure 6.1. The field-independent elements of an argument include a *claim* based on *data* (facts/evidence) connected to each other through *warrants*: the reasoning defending the claim based on the data. 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, rigour or impossibility of arguments or conclusions, we must ask these questions within the limits of a given field…' (Toulmin, 2003, p. 236).

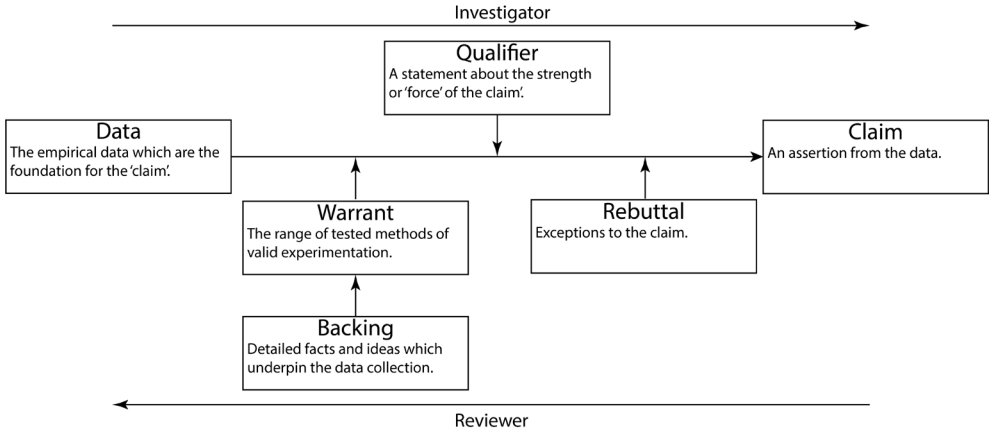


Figure 6.1: Toulmin's argumentation model (Toulmin, 2003) adapted by Gott and Duggan (2007) to secondary science inquiry

A useful tool to think about the field-dependent content of an argument in scientific inquiry is provided by the PACKS model shown in figure 6.2. In this model, Millar et al. (1994) link decisions made in various phases of an inquiry to four types of knowledge involving: (**A**) the purpose of the inquiry, (**B**) the relevant content, (**C**) the required manipulative skills, and (**D**) the quality of scientific evidence. With regard to argumentation PACKS knowledge type **A**, e.g., influences students' interpretation of the task and thus influences the type of claim made by students (Millar, 1997). While each of these knowledge types influence the decisions being made, knowledge type **D** is especially important in the construction of an argument in support of a claim.

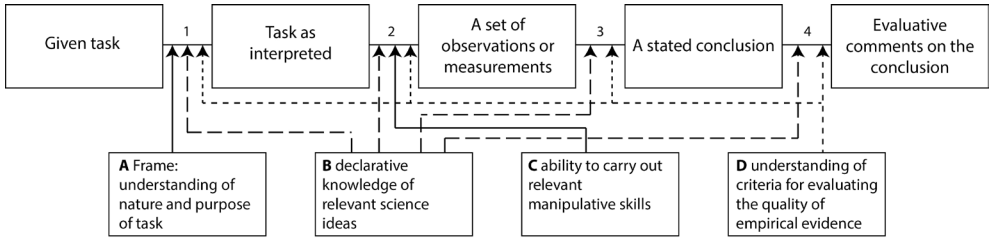


Figure 6.2: The PACKS model identifies various types of knowledge and their influence during inquiry tasks.

Prominent elements in knowledge type **D** are the so-called *Concepts of Evidence* (CoE) (Gott & Duggan, 1996; Millar, 1997). The tentative list of 93 CoE (Gott et al., 2003) contains concepts such as accuracy, range, and interval, which underpin the umbrella concepts validity and reliability of data. These CoE are the building blocks that enable us to construct, analyse, and judge a cogent account of the evidence (Gott & Duggan, 2007; Gott & Roberts, 2008; Pols et al., 2022a). However, a researcher does not conceive and construct a QPI using individual concepts but rather relies on insights in which these individual

concepts acquire meaning through their relation to other concepts (Pols et al., 2022a; White & Gunstone, 1992). Pols et al. (2022a) explicated these insights as so-called *Understandings of Evidence* (UoE) in which the CoE are constitutive elements (see Table 6.1 for examples). In an augmented Delphi study, they identified 19 UoE distributed over six inquiry phases. The study rendered a set of UoE that is validated as necessary and sufficient for evaluating evidence in QPI at secondary school and first year university level. By specifying indicators for various levels, the study provided an 'assessment rubric for physics inquiry' (ARPI) that allows one to identify for each of the UoE which of five levels of understanding a student has attained. An appropriate selection of these UoE form the set of learning goals for the teaching-learning sequence (TLS) studied here, and the corresponding section of ARPI is used to monitor student progress in their ability to successfully engage in QPI.

Table 6.1: An overview of the UoE (Pols et al., 2022a) that are selected as the learning goals for the TLS with various CoE 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 t**hat provide the required **reliability** and **accuracy** of the dataset.
3 Method & Procedure	8	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 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 & 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.

The PACKS model tells us what types of knowledge students need in order to engage successfully in inquiry, noting in particular knowledge of the quality criteria of evidence. We think it would be an advantage for students to know not only *what* these criteria are and *how* they can be satisfied but also, *why* scientists adhere to them. The UoE express understandings of the nature of QPI and its data that provide this explanation. 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 QPI. 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.

6.3 Aims and research questions

Based on the literature and personal professional experience we specify design principles that are meant to strengthen students' critical evaluation of the quality of evidence, reflection on their own approach, and explicit understanding of scientific evidence. We implement these design principles in the activities of a TLS that integrates argumentation in inquiry and investigate the learning outcomes. We explore their contribution to students' argumentation in inquiry. This is meant, in turn, to guide an exploration of the role of argumentation in inquiry learning and our theoretical understanding of that educational process. The corresponding first research question is:

> **RQ1)** *Do the selected design principles yield the expected returns in terms of students' actions and attitudes?*

If the design principles yield the intended effects, the next question is whether students attain the intended learning outcomes and develop targeted understandings that underpin scientific argumentation in inquiry:

> **RQ2)** *To what extent do students attain the targeted UoE during the activities?*

Finally, if students develop these targeted understandings to a certain degree, then how does this affect their argumentation, and does this actually contribute to their ability to set up and conduct a rigorous QPI? Specifically:

> **RQ3)** *In the TLS, what progress is observed in students' ability to engage independently in QPI?*

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 QPI?*

6.4 Method

6.4.1 Research design

In this study we explore the integration of argumentation in inquiry in a way that, while advocated variously in the literature (Driver et al., 2000; Gott & Duggan, 2007; Hofstein, 2017; R. Roberts & Johnson, 2015; Watson et al., 2004) to our knowledge has not been subjected to empirical study. As in educational design research (EDR) (P. Bell, Hoadley, & Linn, 2004; Van den Akker et al., 2006) we are committed to simultaneously develop theoretical insights and practical solutions (Barab & Squire, 2004; McKenney & Reeves, 2013) through a combined study of both the process of learning and the means that support that process (DiSessa & Cobb, 2004; Van den Akker et al., 2006). This approach can potentially '*fill the gap between theory and practice*' and '*lead to interesting results in terms of pragmatic value (feasibility, effectiveness, etc.) and/or results in terms of scientific validity such as understanding learning processes*' (Barab & Squire, 2004; P. L. Lijnse, 1995; Méheut & Psillos, 2004; Van den Akker et al., 2006). In accord with suggestions from literature (Crawford, 2014; Hodson, 1990; Hofstein & Lunetta, 1982; Millar, 2010), the feasibility of the design principles and the development of UoE in the TLS were established in a small-scale single classroom case study that allowed for closely monitoring, evaluating and interpreting students' responses and actions.

For each of the selected design principles we specified the students' actions and attitudes (RQ1) that would be observed in case of effective implementation, and compared these with what happened in each of the five activities. Students developing understanding of evidence (RQ2) was investigated through meta-cognitive tasks. Students expressed what they learned and put that knowledge into practice. We compared these statements with the targeted UoE and established their ability to operationalize the targeted knowledge. Using internal evaluation (Méheut & Psillos, 2004), students' enhanced ability to engage in independent QPI was determined by comparing their understanding of and their attempts to adhere to scientific inquiry criteria (Pols et al., 2022a) at the start and end of the TLS (RQ3). Triangulation of the pre-post comparison with the data obtained during the various activities allowed us to establish which aspects of the teaching were particularly important (Andersson & Bach, 2005; Méheut & Psillos, 2004).

6.4.2 Participants and educational setting

The TLS was carried out by the first author in the first two months of 2019 (see Table 6.2) at a regular Dutch school. It took place in his class of twenty-three students, aged 14-15 and in their final year of lower secondary education (Grade 9), during their regular 50 minute physics lessons. He designed the TLS and had ten years of experience in physics teaching. As a trainer in an in-service professional development course focused on teaching scientific

inquiry (Pols, 2021b) he was familiar in particular with the challenges in QPI.

Students' work, used as data for the study, was graded, but handing in the work sufficed to obtain a pass grade (7 out of 10). The students worked in largely fixed, self-selected teams of two students, or three to assure no-one worked alone. Due to illness, team G2 was cancelled after activity 2 and does not appear in the data after that activity.

In Dutch lower secondary education, physics is mandatory and focusses 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 but no national exam at the end of lower secondary education (Ottevanger et al., 2014; Spek & Rodenboog, 2011), teachers are to a large extent free to devise lessons and teach in the way they deem fit. A more detailed description of the Dutch educational system and the population of our sample is provided in Pols et al. (2021). There we argue that the sample is not exceptional, and findings representative of many similar educational settings.

All names are fictitious, all data were collected and treated in accord with relevant ethical guidelines. All interventions, instruments and collected data were in Dutch and have been translated by the authors. Practical aspects of the TLS are described in more detail in Pols et al. (2019). Materials for all associated activities are open-source available in English, Dutch, French, Spanish and Basque (Pols, 2021a).

Table 6.2: An overview of the activities with the targeted UoE, number of teams participating and the data collected.

Activity	Week	Data sources	Design principles	Targeted UoE	Main learning objective
1. Pirate pendulum	1	i, iv, v	1-3		Developing the notion that in QPI the best available answer is to be produced
	2	ii, iii, iv	4-5		
2. Tricky Tracks	3	iii, iv	1-5		Distinguishing observation from interpretations and raising awareness of the need for argumentation in inquiry
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
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
5. NASA's CRV	6	i, iv, v	1-3	Application of all of the above	Applying the acquired knowledge in an integrated way
	7	i, iv, v			
	8	ii, iii	5		

6.4.3 Educational design

Educational aims and approach

The TLS aims at the development of a deeper understanding of the scientific purpose of QPI and PACKS type **D**, specifically of the UoE that relate to the collection of reliable and valid data and provide a solid basis for developing other UoE, see Table 6.1 and 6.2. Students' focus on both the quality of the data and the purpose of QPI is meant to foster learning of *how* to get reliable data *why* reliable and valid data are important, *and* why data obtained in this way can be regarded to be reliable and valid (Kind, 1999; Osborne, 2014a, 2014b).

Design principles

The following design principles, summarized along with their expected returns in Table 6.3, are meant to support students' critical evaluation of the quality of evidence and reflection on their own approach:

DP1 Guided inquiry This design principle seeks a balance between autonomy and guidance. 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 (Glaesser et al., 2009; Hodson, 2014; Hofstein & Kind, 2012; Holmes & Wieman, 2016, 2018; van den Berg, Buning, & Smits, 1996; Zion & Mendelovici, 2012). Enough guidance is required to overcome insuperable problems that inexperienced students will encounter in inquiry (Hodson, 1990; Johnstone & Wham, 1982). *Guided inquiry*, where the research question is given but students follow their own path to construct an answer (Tamir, 1991), is expected to offer this balance for the novice population at hand.

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 Cognitive load, creating barriers to learning, is reduced by avoiding distracting details pertaining to PACKS type **A-C** (Hart, Mulhall, Berry, Loughran, & Gunstone, 2000; Hodson, 1988, 1990, 1994, 2014; Jenkins, 1998; Johnstone & Wham, 1982; Tasker & Freyberg, 1985; van den Berg, 2013). To do so, we ensure that the activities are easily understood (**A**), as simple as possible in terms of equipment (**C**) and of required conceptual knowledge (**B**). Students' questions related to these PACKS types are meant to be answered straightforwardly by the teacher so that the focus is on developing type **D** knowledge.

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. If it is satisfied students focus, during and subsequent to answering it, on the quality of the answer and of the evidence supporting it.

DP3 A real life context Implementation of research findings can easily be shown and understood to affect e.g. people's safety. Consideration of a real life, meaningful context (Gilbert, 2006; Kortland, 2007; Ntombela, 1999) therefore makes it easier to see why trustworthy data and conclusions are needed. This understanding of the *practicality of consequences*, number 87 in the list of CoE (Gott & Duggan, 2003), can help students understand why finding a scientifically convincing answer is relevant. In turn, it might provide the motivation to invest the effort required to find such an answer (Pols et al., 2021, 2022b) and a need to extend their scientific knowledge (Kortland, 2007). However, it is understood that merely contextualizing the problem does not necessarily enhance students' critical approach. Students may easily see the context as window-dressing or simply forget it (Molyneux-Hodgson et al., 1999; Pols et al., 2021, 2022b). 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 a useful answer to their research question from the context framing the research problem. It is satisfied if students are seen to regard answering the research question to be relevant and worthwhile.

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 (Barron et al., 1998; Gunstone & Champagne, 1990) where intervention by and negotiation with the teacher are essential (Driver, 1995). As in other educational activities (Kapur, 2008, 2011; Roll, Holmes, Day, & Bonn, 2012), we make productive use of decisions that students upon reflection regard to be sub-optimal. 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*' (Millar, 2009). 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 if students' methodological decisions in inquiry are observed to become the centre of their attention in the activity. It is satisfied if students actively engage in becoming aware of their decisions, in evaluating these, in identifying their strong and weak aspects, and in using these insights to direct their (future) decisions.

DP5 Meta-cognitive 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. Students' metacognitive awareness of their understanding (Dehn, 2011), important in successful learning (Livingston, 2003), is to be consolidated in meta-cognitive tasks that invite them to reflect on, value and organize the targeted knowledge (Larkin & Reif, 1979). These metacognitive tasks consist of the following two sentences to be completed by the students in each activity:

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

As metacognition is often defined as "thinking about thinking" (Livingston, 2003), these questions can thus be considered metacognitive tasks. Any additional meta-cognitive tasks are specified in the description of the activities.

This design principle is satisfied if students are observed to actively engage in reflecting on their newly obtained insights about establishing, critically evaluating and defending evidence in support of a claim that answers a research question. Evidence as to whether this principle is satisfied is to be found in the words and actions students use to express, value and organize that knowledge.

Table 6.3: The design principles and features derived from literature and the rationales, and expected returns

No.	Label	Rationale	Expected returns
1	Guided inquiry	Offers balance between autonomy and guidance	Students make their own methodological decision, 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 the given 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	Meta-cognitive 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

The activities

Activity 1: The pirate pendulum

After viewing a film clip of pirates swinging from ropes between sailing ships (Bruckheimer, 2007), students investigate a model, the pendulum (**DP2**), to help the stunt coordinator produce a safe but spectacular film stunt (**DP3**). Students set up their own inquiry without further instructions (**DP1**). After submission of their reports, students are asked: 'if the stunt was based on the information you have provided, and if you were the stunt(wo)man, would you feel safe to jump?'. Students are meant to take a different perspective, identify shortcomings in their work (**DP4**) and formulate personal criteria for evidence of acceptable quality (**DP5**).

This inquiry is a first step in developing the notion that answers to research questions are useful only if they are trustworthy and optimally informative. From students' own ideas on what is needed for the stunt to be performed safely, they specify their own 'scientific' standards.

In activity 1 students are meant to come to understand that in scientific inquiry, only the best possible answer, given the circumstances, is good enough. In the remainder of the TLS, they take first steps in finding out what this 'best possible answer' means in science. The latter is the subject of this paper while the specific role of activity 1, its aims, outcomes and the baseline inquiry knowledge of students have extensively been described elsewhere (Pols et al., 2022b). The results section pertaining to this activity is therefore restricted to verifying realisation of the design principles and establishing students' baseline ability to engage independently in inquiry.

Activity 2: Tricky tracks and the existence of the Yeti

Activity 2 takes the first step in exploring what 'the best answer' entails in science. It does so by targeting students' ability to distinguish between observation and interpretation of data using an adapted version of Lederman's 'Tricky tracks' (1998). Students observe a picture on the interactive whiteboard and are asked to express, taking turns, something they observe in it without repeating each other's earlier statements (**DP2**). Given this particular picture, someone is bound to say something like '*these are tracks of two birds in snow [or sand]*'. Once all students have made a statement, the teacher discusses whether students' statements actually express *observations* (**DP4**), i.e., things that can simply be seen (or detected with the senses) and therefore agreed upon. Students are meant to discover that many of their statements express interpretations rather than observations. They are to note that several interpretations of a single data set may exist that are potentially equally valid and that these interpretations may therefore be contested.

Just as in actual research students are encouraged to construct a convincing argument by providing reasons (*warrants*) supporting their own interpretation of the data

and invalidating alternative interpretations. They experience that this is possible to some extent, but never fully. It is meant to serve as an exemplary situation in which the importance of justifying one's interpretation of data is of the utmost importance.

A basic form of Toulmin's argumentation model (Toulmin, 2003) is presented as a first step in constructing the conclusion to an investigation as a scientifically convincing argument in support of a claim. Students consolidate their gained understanding by analysing a newspaper article (Phys.org, 2011) in which '*irrefutable evidence*' about the existence of the Yeti is claimed (**DP3**) and assess this claim. Students' ability to distinguish between observation and inference is tested by asking for their written observations of a picture of a female basketball match at the Olympics.

Activity 3: Investigating the 'ape index' for the International Swimming League

The fictitious International Swimming League (ISL) is said to consider length classes in competitive swimming, as people with relatively long arms may have an unfair advantage (Lavoie & Montpetit, 1986) (**DP3**). Students contribute to the ISL investigation by taking measurements of classmates using a tapeline and determining the ratio in humans of body length to arm span, often referred to as the 'ape index' (**DP1 & DP2**). No guidance is given on how to measure. Once the students have shared the data on the interactive whiteboard the reliability of the data is discussed (**DP4**). Subsequently, the validity of the investigation is discussed, where we expect students to consider that a very specific sample of students aged 14-15 from a specific region in the Netherlands (UoE 9, 16) was used. Next, students' measurements are combined with a larger data set that reveals a pattern, and a line of best fit is provided. Jointly, students evaluate several conclusions about the pattern, presented by the teacher in order of increasing precision and detail. They discuss '*What conclusion is of most value to the ISL, and why?*' (**DP3**). Using the scaffold notion that a conclusion is to be as trustworthy, useful and informative as possible, students are asked to write an even better conclusion to be submitted to the ISL. They are then asked what conditions a conclusion must meet in order to be a good conclusion. Students are asked to write down whether they would like to change their conclusion to the earlier 'Pirates' activity, and if so, how.

The activity fosters further development of the understanding that a conclusion should be optimally trustworthy, useful and informative. Students explore ways to obtain that kind of conclusion. The activity augments the previous activity by exploring the structure and content of an argument (*backings*). It focuses on the features of the data set (reliability of the used method, measured range, sample size) and how these contribute to, or limit, the reliability and validity of the conclusion (*qualifiers*).

Activity 4: Car crash barriers

A cogent claim is built on reliable data. As variability in measurements is inevitable, finding

an estimate of the 'true' value requires repeated measurements. Students are to develop this notion, and the understanding that a claim stating a relation between quantities becomes more credible if it accounts for the variability (providing *backing*).

To develop these insights, students investigate the influence of the strength of crash barriers on the stopping distance of a car. Selecting the strength of these barriers for a slippery road near a canyon is delicate; too weak and the driver will end up in the canyon, too strong and the driver will be harmed by the impact of the crash (**DP3**). The situation is modelled by a marble rolling from an incline crashing into an inverted pile of paper or plastic cups (**DP2**) (Farmer, 2012). Repeated readings clearly show natural variability in the stopping distance. Students select their own materials, procedure, ramp steepness and marble's starting position (**DP1**).

During the activity the teacher challenges students gently, as in: 'your results keep on changing, are you sure you are doing a proper job?'. They are to note that the variability in repeated readings is unavoidable (**DP4**). Once the graph has been produced, the results are compared and discussed. The teacher is meant to explain that variability in the data is natural and that more variability makes findings less trustworthy (UoE 8). Students ought to see the sense of repeating measurements and of reporting not just the average but also the spread in data (**DP5**). Subsequently two concept cartoons (Keogh & Naylor, 1999) are discussed in which students consider different actions on a set of repeated readings, see Figure 6.3. Students are asked to write down what they would change to their procedure used in the earlier 'Pirates' activity, activity 1. Finally, students draw conclusions based on their acquired data after being reminded of their establishing in activity 3 that *conclusions ought to: (1) answer the question, (2) be reliable, (3) be as informative as possible, and (4) be useful*.

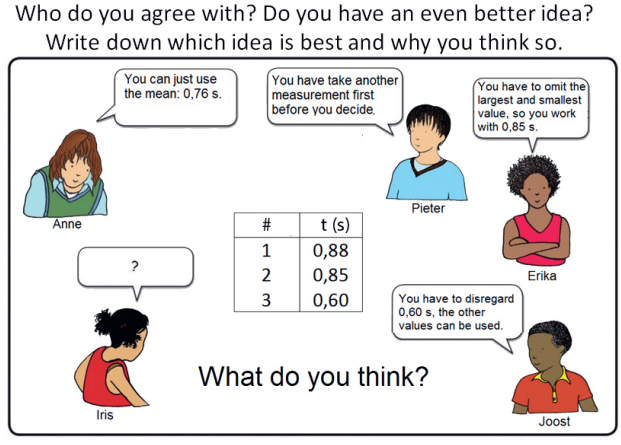


Figure 6.3. One of the two concept cartoons used to have students grapple with the ideas of variability and repeated measurements

Activity 5: Investigating NASA Crew Return Vehicle

Where activities 2-4 aimed at understandings that underpin optimally cogent answers in science, this final activity is meant to consolidate previous learning and monitor the independent application of these UoE. Students conduct a QPI in which they determine the relation between the falling speed of a Crew Return Vehicle (CRV) (Loren, 1992) – modelled by a paper cone (**DP2**) (Mooldijk & Savelsbergh, 2000; Mooldijk, van der Valk, & Wooning, 2006) – and one of the factors that potentially influence it (choosing e.g. mass, shape, frontal area, …). The task description states that this information is required for a highly advanced computer model to help NASA in developing a new CRV that safely but quickly returns astronauts home (**DP3**). Students investigate the relationship based on their own research design (**DP1**), where they are reminded of their 'rules for doing proper investigations'. As part of reporting their outcomes to NASA, teams review each other's reports, provide feedback and process it. Finally as a meta-cognitive task students write their *main tips for doing QPI well* as a letter of advice to next year's students (**DP5**). Design principle 4 is not implemented as the activity focuses on establishing the acquired knowledge rather than developing new insights.

6.4.4 Data sources

Students compiled a portfolio to help them evaluate what was done and learned (Partnership). For research purposes, these portfolios were handed in after the final activity. The portfolios contained:

i) Scientific Graphic Organizers (SGO) providing data for RQ1 and RQ3 are used in activities 1 and 5. An SGO is a pre-structured lab journal where students report the essentials of an inquiry without the necessity to write an extensive lab report (Pols, 2019; Struble, 2007). Elements of the SGO are research question, methods, instruments, essential theory, data represented in table and graph, and conclusions. 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 SGO. 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 meta-cognitive tasks of each activity in which students express their perceived learning gains which contribute to answering RQ2 about the students' developing UoE. Students' UoE are also elicited through 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** giving a summary of each activity.

6.4.5 Data analysis

RQ1: Successful implementation of the design principles

Whether DP1-2 were successfully implemented was established by verifying whether students produced the intended output (data, graphs, answer to the question). On the basis of audio recordings 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 analysed whether addressing the weaknesses in students' approaches (DP4) triggered discussions in which the issue at hand became the centre of attention. For DP5 we established whether students produced solid answers in the reflective task and used these self-perceived insights to forward points of improvement pertaining to the 'Pirates' activity.

RQ2: Attainment of UoE

For each activity we verified whether the students' perceived learning gains concurred 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 and that the *minimum* level of attainment of an UoE is derived from the student's *actions* and *justifications* (Pols et al., 2022a). If a student makes a scientifically acceptable decision, e.g. repeats a measurement and reports the mean, the intermediate level (level 2, see Table 6.4) 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 (Giddings et al., 1991), we run the risk of underestimating a student's understanding. If a justification is lacking does not mean a student is unable to give it. While it has limitations, ARPI does provide a means to tentatively derive students' understanding from their actions.

Table 6.4: 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 not fully reached.

UoE	The researcher understands that:	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. Does not consider collecting further data at any stage.	Repeats measurements a fixed but sufficient number of times without substantiation in terms of the quality of the dataset. Considers collecting additional data only in retrospect, as a recommendation.	Substantiates the required number of repeated measurements based on the spread in the data and the required reliability. Considers collecting alternative, additional data and collects these if appropriate.
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.

RQ3: Students' ability to engage in QPI

We explored and described the development of students' ability to engage in inquiry independently in terms of a selection of indicators that were addressed specifically in the activities of the TLS. We compared activities 1 and 5 in terms of whether students spontaneously:

a) Construct the inquiry as an argument for a claim
b) Take variability 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
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
f) Apply a critical attitude towards their own approach and findings

Criteria *a*, *e* & *f* relate to understanding the scientific purpose of QPI that motivate finding the best available answer to the research question. Criteria *b-d* pertain to the basic choices that in every QPI ought to be made, but are rarely adequately taken by students at this age. This set of criteria are therefore indispensable, yet tentative, in conducting a QPI independently. In our view, it is the very first basis that students need to be able to further develop knowledge and skills with regard to argumentation in inquiry. If students spontaneously meet criteria *a – f* they are applying in their inquiry the understandings and attitudes we aim to develop in them.

The analysis is based on the students' actions, decisions and justifications reported in their SGO and written report. A qualitative comparison in activity 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. A further, in-depth qualitative analysis of classroom decision-making discussions among students and of consultations between students and teacher reveals some of the thinking behind the doing. It enables us to evaluate how the students' approach towards inquiry changed over time. However note that we analysed the students' ability to engage independently in inquiry on the basis of the PACKS that can be expected of novice science students with virtually no experience in inquiry.

6.4.6 Reliability and validity

Studying one's own educational practice has the potential to bridge the research-practice gap (Vanderlinde & Braak, 2010), has high ecological validity (Bryman, 2015), is accepted in

both action research (Altricher et al., 2005) and educational design research (McKenney & Reeves, 2013; Van den Akker et al., 2006) and advocated especially in the area of scientific inquiry (Crawford, 2014; Hodson, 1990; Hofstein, 2017). Potential threats to data analysis bias (Mason, 2002; Trowler, 2011) were minimized. The main data analysis (application of ARPI) and a significant part (30%) of secondary data analysis (student work) was carried out independently by the first author and a second teacher/researcher conversant with the teaching sequence. Rare cases of disagreement were discussed until 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 involves 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 team work and whole-class activities. Strictly speaking, we therefore cannot regard these data as reflecting *individual* attainment levels. We consider this to be an acceptable trade-off, since individual work would have overly affected the authenticity of the lessons, where team work 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.

6.5 Results

The research questions ask us to explore how implementing the design features of the activities contributes to students' ability to conduct independent inquiry. In presenting the data, we therefore first explore whether the actual classroom activities have the characteristics that DP 1-4 are meant to effect (RQ1). We also illustrate the resulting learning process with brief vignettes. We then study whether students reflect on their new insights (DP5). We describe and interpret their expressions in terms of their alignment with selected UoE (RQ2). We present the data in the order they were collected where possible. The data on activity 5 are used finally to establish the overall, integrated progress in argumentation and its influence on the quality of students' inquiry (RQ3). To better compare students' progress pertaining their ability to engage in QPI, data of activity 1 for each of the six criteria are given along with the data of activity 5.

6.5.1 Activity 1: The pirate pendulum

Design principles – realization and learning process

DP1 Guided inquiry & DP2 Reduction of knowledge demand

After watching the movie scene and receiving the assignment, the students went to work with enthusiasm, planning and conducting the experiment. Their actions were aimed at answering a research question they understood and valued. Since all students produced appropriate data with suitable procedures that allowed them to establish a relation between two variables, DP1&2 were implemented successfully. Further details are provided in Pols et al. (2022b).

DP3 Real-life context

Students' enthusiasm, drive, and reference during their work to the context indicated that overall DP3 had been implemented adequately. In the first stage of the activity, the answer's quality, scientific or otherwise, seemed to have no relevance for students. This however agreed with our predictions and as expected, students' views and attitudes changed dramatically when they faced the question: if we plan the stunt according to your results, would you dare to jump? The context appeared to be instrumental in making the practicality of consequences relevant to students, and raising their awareness of the importance of the quality of scientific evidence.

DP4 Productive failure & DP5 Meta-cognitive tasks

Once the students decided that their results were not good enough to guarantee their own safety, their methodological decisions became the centre of their attention as is seen in an exemplary statement of Team G8:

> G8: We could have used a camera that allows to measure the time accurately, and a setup that allows to repeat the exact procedure each time.

While this illustrates that DP4 was satisfied, students also engaged actively in reflecting on what they had learned in these activities, confirming implementation of DP5:

> G2: Use the same procedure each time and carry out the same test multiple times.
> G4: We must be critical of the conclusion we draw.
> G9: Research has to be conducted thoroughly as otherwise the results are not reliable. Furthermore, it is important to draw clear conclusions as they are not useful otherwise.

We see that DP4&5 were implemented successfully in that students critically evaluated the quality of their approach. Students' self-perceived learning outcomes align

with the intended outcomes: they developed the notion that an informative conclusion is needed and requires a rigorous experiment. The activity resulted in a readiness for learning the insights required to devise an experiment of that kind, *i.e.*, for learning the targeted UoE.

6.5.2 Activity 2: Tricky tracks and the existence of the Yeti

Design principles – realization and learning process

DP1 Guided inquiry & DP2 Reduction of knowledge demand

When the 'Tricky tracks' picture was displayed Julia (G11), an engaged student of average academic ability, was the first to express an observation about it:

Julia:	It concerns two birds.
Thim (G10):	Birds?
Julia:	Yes.
Thim:	Huh? How?

The subsequent statement, '*footsteps*', evoked a response from Julia:

Julia:	Footsteps? Ah, that is also possible.

Successive 'observations' were subjected to scrutiny, where students asked each other for clarification. So far, the teacher had not spoken.

DP1&2 were implemented successfully as students understood the question 'what do you observe?' and required no other prior knowledge. As we expected, they interpreted the term 'observation' informally rather than scientifically. The spontaneity of students' focus on the issues of observation, interpretation and their role in evidence surprised us. It is indicative of the educational strength of Lederman's 'Tricky Tracks'.

DP3 Real-life context

The claim of proof of the Yeti's existence in a scientific approach ought to be evaluated in terms of the quality of the evidence rather than of preconceived opinion. Students showed they were able to do so:

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

Three minutes later:

Julia:	I just don't believe it. Here it says they. But who are they and can we trust them? It also states that they want to use the Yeti to attract tourists, it might just be that they made up the story.

Teacher: So, do you consider it to be evidence?

Julia: No. […] I would at least film the Yeti and study its behavior for at least a week.

Students' worksheets displayed similar but sometimes inferior critiques on the quality of the evidence. For example, four of the eleven teams used a circular argument: 'The evidence cannot be used for the claim because it is not certain that it came from a yeti.'

Students took the 'real life' contexts seriously, and using these had a positive effect on their cognitive and emotional engagement with the issues at hand. The approach was successful in eliciting relevant ideas that, though informal, appear to align with scientific notions of what constitutes evidence.

DP4 Productive failure & DP5 Meta-cognitive tasks

The teacher in Tricky Tracks asked whether the 'observation' that '*it concerns birds*' could be contested. Pointing at contradictory statements, the students agreed that *all* entries could in fact be disputed and should therefore not be called 'observations'. What observations are, and what can be inferred from these was further investigated by exploring what 'evidence' one would look for if the actual location could be visited. Several ideas were put forward (feathers, poop, rests of food), and the interpretation of this evidence further discussed (e.g. can we be sure that two birds were present at the same time?).

We see that an exploration of the statement 'it concerns two birds' as a carefully chosen 'bad example' appeared to instigate further student thinking and the construction of new, useful insights about the meaning of concepts such as 'observation' and ' inference', about the distinction between these concepts, and about the problem of interpreting observation as evidence in support of a claim.

The data presented in the next section show that DP5 was implemented successfully: students engaged thoroughly in reflecting on what they had learned.

Development of UoE

Students' written work was explored for evidence of students' ability to distinguish between observation, inferences and conclusions. While not explicitly identified as a separate UoE, this insight is part of the development of a method for constructing and evaluating cogent conclusions. The most comprehensive rules for doing proper investigations were formulated by teams G8 and G10:

G8: An observation is what you see. A conclusion should state its likeliness if there are multiple possibilities.

G10: Always first observe, then critically think whether the observations are correct. You then state several conclusions and compare these with

what you observed. You have to use the facts rather than use what you think happened.

While naïve, these statements do show that students were reflecting on the concepts they used in producing evidence. Students' reflections on what was learned accorded with the intended learning outcomes:

G3: We learned that there is a difference between observations and conclusions. We learned how to substantiate a conclusion.
G7: You have to check whether the source is reliable, you have to make good observations and not draw conclusions if you are not really sure.

The reflective task showed that the students evaluated the cogency of conclusions by expressing appropriate criticism on both the quality of the evidence and the way it was obtained.

G7: The hairs and the footsteps can belong to another animal. The territory and the sleeping place can be made by men, or have been made up.

However, students were not yet able to fully apply their acquired insights in the final task. For example, they failed to see that their 'observation' of '*a female basketball match at the Olympics*' involves an interpretation of the symbol of five rings at the side of the field.

Students' responses align with the targeted scientific ideas, but are insufficiently complete and detailed to be operationalized in subsequent activities. Not surprisingly, a first step was made but practice in more diverse situations is needed.

6.5.3 Activity 3: ISL

Design principles – realization and learning process

DP1 Guided inquiry & DP2 Reduction of knowledge demand

Students measured each other's body length and arm span to examine a relationship between these variables. A few teams did so by standing straight against the wall, but most did not. Only one student took off the shoes. Various teams were observed to measure arm span with arms not fully stretched. Students collected the data in a mere seven minutes using the provided tape measure, none asked the teacher for help.

We see that DP1&2 were implemented successfully in that students collected the necessary data following their own methods without impediment. However, limiting the knowledge demand did not result in all students focusing on devising a reliable data collection method. Most teams collected the data rapidly and seemingly unthinkingly.

DP3 Real-life context

The context was considered by the students only after data collection, when the teacher

imposed it, criticizing their research methods:

Teacher:	Can these data be used by the ISL? Has each one of you measured in the same way?
Masha (G5):	Probably not.
Teacher:	Probably not... Who took off the shoes? [one hand is raised] Only one of you took off their shoes. Do your shoes count as body length?
Students:	No.

Subsequently the teacher presented several obviously inadequate poses he had seen students use to measure lengths, asking for comments. The teacher briefly noted that the reliability of data is influenced by how the data are obtained.

The context was used again by the teacher when he asked whether the statement 'taller people have longer arms' was valuable. Various students mumbled 'no', then:

Xander (G7):	It misses an argument for the conclusion.
Tom (G10):	I don't think it is a proper conclusion for the ISL, it doesn't reflect the ratio.
Teacher:	[later inviting comments on:] 'There is a proportional relation between length and arm span.'
Thim (G10):	Yes, but there are still exceptions.
Teacher:	Good, so you would mention that as well.

The teacher presented and asked the students to reflect on further conclusions of increasing quality. He ended by summarizing what characterizes that quality: conclusions should answer the question, and to be satisfactory to the ISL, should be as trustworthy, useful and informative as possible.

The ISL context in itself did not instigate most students to invest effort in doing quality work. However, it was successfully exploited in discussing the value of the various conclusions, making the elements that constitute a proper conclusions more tangible to students. It seems to help Tom, for example, to realize that a quantitative conclusion would be preferable.

DP4 Productive failure & DP5 Meta-cognitive tasks

Students' choices in measuring body length and arm span were evaluated and when in retrospect deemed inadequate by them, as in the cases of not taking off shoes or standing straight, utilized as 'bad examples'. Discussing the utility of the students' data the teacher highlighted the importance of choosing suitable procedures to get valid data. This was found (see below) to contribute to students' attainment of UoE 6.

Students actively engaged in evaluating their new knowledge, the intentions of DP5 were realized. Illustrative examples of their engagement demonstrate this in the next section.

Development of UoE

Students' reflection on what they learned reveals progress in attainment of UoE 6 and UoE 13. S*ix of ten team responses related to the targeted UoE 6* (G3-6,10,11). Some examples:

> G5: You should measure each person using the same procedure (for instance, each person has to take off their shoes), you have to measure accurately, be clear what you are talking about.
> G6: Measurements should be collected in the same manner to be reliable.

Eight responses (G1,3,4,6-10) referred to conclusions, as seen in these examples:

> G3: It is difficult to draw a proper conclusion, it should be based on facts and not on an opinion to be reliable.
> 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.

Six teams showed awareness that the inquiry should result in an informative conclusion, e.g., in proposing improvements for the earlier 'Pirates' activity:

> G9: We should be more precise with a more comprehensive conclusion in which everything is described. The results would then become more useful.

Several teams referred to providing justification to back up conclusions:

> G1: We learned how to formulate a proper conclusion, for instance that it should be verifiable and clear. The research results must also be included in the answer.
> G11: [A proper conclusion provides:] An answer to the research question with proper arguments that you were able to demonstrate in the experiment.

However, in drawing informative conclusions students still encountered several difficulties, e.g., in interpreting data:

Teacher:	What conclusion do you present to the ISL?
Julia:	[remains silent]
Teacher:	Is there a ratio?

Julia: Not really. I see that taller people have longer arms, but that is not useful.

Students' responses aligned with the intended learning outcomes, but they were unable to specify these insights at a level of abstraction that would be needed for transfer to future inquiries. The developing attainment of UoE 6, about selecting optimal measuring methods, was not explored further after this activity. We infer that students appeared to understand that an informative conclusion is one that is substantiated by evidence (UoE 13) but that they were not yet able to actually construct that kind of conclusion.

6.5.4 Activity 4: Car crash barriers

Design principles – realization and learning process

DP1 Guided inquiry & DP2 Reduction of knowledge demand

Unimpeded, students produced the intended graphs. In discussing the results, they recognized that – as result of slightly different setups – the graphs were similar in shape but representative of different measurements:

Noah (G7): Our results are not the same as we have a different marble and a different angle.

They were able to relate the experimental features mentioned by other students (kind of cup, starting height, etc.) to the car's mass, velocity and even the slipperiness of the road.

DP1&2 were implemented successfully in terms of students producing the desired graph while their prior knowledge sufficed. The simple model of the inverted cup was easily related by the students to the actual context of the investigation. Since placing marks at the final locations of the cups immediately results in the data points of the graph, the data are immediately interpretable, further limiting the knowledge demand. Consequently the focus was on comparison and interpretation of the data.

DP3 Real-life context

The context was frequently used by the teacher to discuss whether students measured properly:

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 to collect some more. Without being asked to do so, 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.

We find DP3 effectively implemented as students were motivated to obtain convincing evidence, albeit not always without teacher's interference. G1's action – spontaneously taking responsibility for the lack of quality of the data by repeating all measurements – was exceptional but indicative of the potential of utilizing a context in holding students accountable for the quality of their work. Note that the approach does not prevent issues related to point and set reasoning (Lubben et al., 2001), but does make them accessible to teaching. This, however, is beyond the scope of this study.

DP4 Productive failure & DP5 Meta-cognitive tasks

The two examples above show that the variability in students' measurements was at the center of their attention during the teacher-initiated talks. During these discussions, students expressed that they did not understand why measurements were not the same, since they used the same procedure each time:

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 unavoidably large spread in measured values, seen as 'bad' by the students especially after the teacher's prompting, encouraged reflection on the measurement procedures. It was also used to question the value of the data:

Teacher:	If we look at this graph, what is the value that corresponds with one cup? Is there a true value?
Eva (G5):	No, the measurements are quite far apart.
Teacher:	What is then the use of repeating measurements?
Thim (G10):	A more reliable result.
Teacher:	But why does it become a more reliable result?

As students did not succeed in answering the question, the teacher explained that measurements inevitably deviate and that repeating measurement provide at least a sense of how well a value could be determined. He concluded, using Thim's statement, reporting the average value alone does not suffice as the car could end up in the canyon.

The various discussions show that the unavoidably large spread in measured values, seen as 'bad' by the students themselves, encouraged reflection on the measurement procedures, implying that DP4 was effectively implemented. Below, examples show that this is the case for DP5 as well. Students suggested improvements for their previous approach in the 'Pirates' activity based on their newly acquired insights.

Development of UoE

Students' learning is reflected in the rules they formulated regarding the number of repeated measured values, reporting the values, or both:

> 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.
> G8: Use all measurements, even the outliers. Three is the least of required repeated measurements, otherwise your findings will be unreliable. If you have good equipment then repeating measurements is not required.

The concept cartoon showed – for nine out of ten teams – that most considered three repeated readings to be insufficient. Team G11 refrained from giving an answer because, they said, they did not know what was measured and to what purpose. Their expressed concerns of the *practicality of consequences* impacted the ideas of other teams as in the second concept cartoon more teams stated that the context was missing. This idea was also forwarded in the formulated rules:

> G7: It is important to consider how and what is measured. Only then you can judge what results to use.

In providing recommendations for improvement, all teams mentioned they would increase the number of repeated measurements taken in the first activity. Five out of ten teams added that they would check for outliers and not only report the average value. These findings suggest an increased understanding of repeatability (UoE 8).

Interestingly, student answers to the final task included not only their conclusions, but also attempts at justification and explanation of how it was obtained. They made use of their current interpretations of the targeted UoE in doing so:

G3: The stronger a crash barrier, the shorter the stopping distance: If the crash barrier is four times stronger, the stopping distance is roughly halved. You have to do additional measurements with 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 five 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 try to provide the argument.

6.5.5 Activity 5: NASA's CRV

Design principles - realization

DP1 Guided inquiry & DP2 Reduction of knowledge demand

Students understood the research question and were motivated to answer it, as evidenced by the SGO's and their engagement in class. The SGO's and fieldnotes also show that all were able to devise outlines of suitable method for answering it. Assistance was asked and provided with regard to the use of accurate instruments. An analysis of students' reports showed that all teams collected relevant data but most experienced problems with the identification and quantization of the observed data patterns.

DP1 is satisfied in that all students showed an attitude towards obtaining optimally convincing evidence. DP2 was partly satisfied. Either independently or with the teachers' help students proceeded well with data collection, but support with interpreting the data, which was carried out at home, was not immediately available.

DP3 Real-life context

Each team reported in a letter to NASA, the fictitious commissioner of the investigation. All teams but one spontaneously provided conclusions with substantiation, allowing for a rudimentary verification of the findings, see the excerpts and conclusions below. Moreover, four teams offered recommendations in their letter to NASA, explicitly mentioning the

inquiry's context.

We see that DP3 was successfully implemented as students invested time and energy in devising methods they saw as reliable (see also below) thereby attempting to produce a useful answer to the research question. We cannot be certain that the context motivated them to do so but note that no grading was used to persuade them.

DP4 Productive failure & DP5 Meta-cognitive tasks

Since activity 5 is meant to establish progress in students' ability to engage in inquiry independently, the teacher's input was minimized, which precludes the implementation of DP4. The following excerpts are taken from the 'letters of advice' students wrote after the final activity 'for next year's students'. They show 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.

G3: Measurements should be repeated so that the conclusion is more certain. Do not rush as it will result in mistakes with the consequence that your results are incorrect and you cannot the answer the question correctly.

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.

Development of UoE

The quantitative data obtained by applying ARPI to the SGO and reports serve to identify salient patterns, providing a global view of development, see Table 6.5. It suggests an improved average attainment in UoE 6 & 8, suggesting a potential progress in students' understanding. However, while some teams consistently attained high levels, others still scored low mostly because explanations and justifications of their choices were absent or brief. While the average attainment of UoE 9 hardly changed, all students spontaneously used appropriate range and interval. However here too they failed to explain why they did what they did. The quantitative data show no changes in average attainment of UoE 14 and 16. Despite an enhanced understanding of UoE 14 apparent in the qualitative data (below), their difficulty with analysing the data resulted in partly unsubstantiated conclusions and ARPI's lowest level.

Table 6.5: Students' attainment levels for activities 1 (Pre) and 5 (Post). Shown is the number of teams per competence level for each UoE (Pols et al., 2022a) on a 5-point scale from lowest (0) to highest (4), on average in SGO and letter. Class average level in grey, deviations larger than 0.5 in the mean score are darker grey. 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

Students' ability to engage in inquiry independently

Students' ability to engage in inquiry independently is established by comparing activities 1 and 5 for each of the six criteria, a.-f., established in the section on data analysis.

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

Activity 1: At the start of the TLS all teams wrote in their letter to the stunt coordinator what was done. For instance, team G6 – an academically average team – reported that they investigated whether the type of rope influenced the swing time (figure 6.4). Only four teams (G1, 7, 9 & 11) reported also how they investigated their research question, while nine out of the eleven letters did not provide the information needed to verify their claim. It was as if data and claim were seen as uncontestable and producing a convincing argument as therefore unnecessary.

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 rope [red. 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

Figure 6.4: Team G6's letter to the stunt coordinator, in bold our analysis pertaining the elements of the letter

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^o, 30^o, 60^o, 90^o 120^o, 150^o). 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.

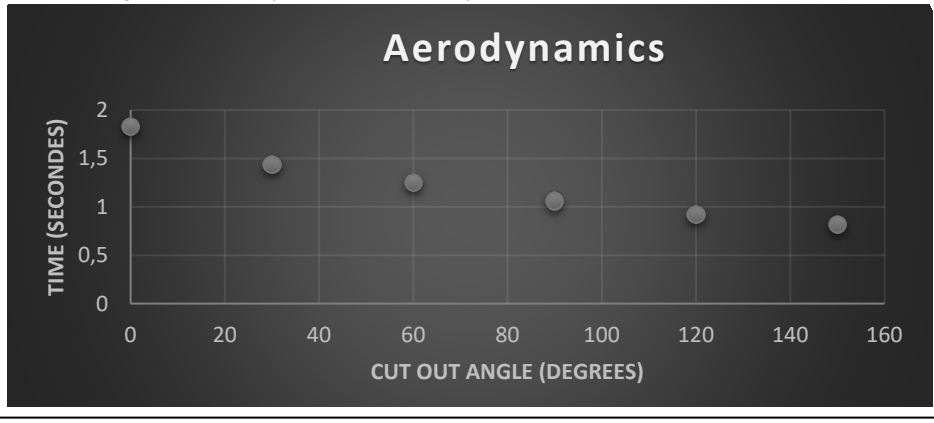


Figure 6.5: Team G6's letter to NASA, in bold our analysis pertaining the elements of the letter

Activity 5: As in the letter of G6 presented in figure 6.5, all teams provided information what was investigated and how it was investigated at the end of the TLS. Nine of the ten teams provided the details, including backings (G1, 3, 5, 6, 8, 9, 11) that allow for an external assessment of the inquiry's quality, shown below. Since these students presented their measurements as an explicit albeit limited substantiation of their answers to their research questions, the inquiry was constructed as an argument in support of a claim.

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

Activity 1: Most teams repeated measurements routinely three times. Since this is generally accepted in secondary school physics but no substantiation of the choice was given the intermediate ARP level was assigned in these cases (see Table 6.5). In their letters only teams G7, 10 & 11 reported that they repeated their measurements. None of the teams was found to consider either the spread in measurements or outliers.

Activity 5: All teams chose to repeat each measurement, usually five times. Two teams (G1 & 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 only the average value displayed in the graph. Five teams linked the number of repeated measurements to the reliability of results, but not with great clarity:

> G6: We have dropped the cone several times (5x) 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 the feedback from others, teams that repeated three 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 seen in Table 6.5 between activities 1 and 5 shows an enhanced awareness of the importance of repeated measurements and how to report them. These qualitative data confirm that trend, but also show that students are not yet able to choose and justify the number of repeated measurements on the basis of the spread in the data.

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 did not optimize their procedure by recording the time for a few swings back and forth. Students were not concerned with choosing suitable instruments and procedures, nor did they consult the teacher about it.

Activity 5: Awareness of the need to produce accurate and precise data through optimizing the choice of instruments and procedures was at this time shown by various teams. Most marked the height on the wall from where to drop cones, to control and measure fall distance. Half of the teams put an effort into optimizing time measurements, e.g. filming the falling cone together with a stopwatch, and analysing the images in slow motion. Two of these teams provided a reasoned substantiation for their choices, G1 doing so 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.

Five teams did not change their approach to measuring time from activity one. Four of these provided recommendations to improve the data collection, but only in hindsight, so ARPI score 1 was assigned. While we observe an enhanced awareness of the necessity to choose appropriate instruments and devise proper procedures, it is not always implemented at the appropriate time, in designing the research.

d. Make deliberate choices in data range and interval

Activity 1: For about half of the teams, there were issues with data range and interval, as they chose only three values or used only a small part of the available rope length. 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.

The variation in approaches suggests that there is no shared understanding of the importance of choosing an optimum range and interval

Activity 5: 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 other teams were assigned intermediate level. While the changes in their unprompted choices suggest that their these are made more deliberately after the TLS, that deliberation is not translated into an explicit justification.

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: Except for team G11 failed effort shown above no other team even attempted to find a quantitative relation and stated a qualitative conclusion. It is noteworthy though that several investigated a relation that according to theory does not exist, and found confirming data. 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.

Only team G3 provided their measurements in a table, no other group used a table or graph.

Activity 5: Seven teams provided a conclusion in terms of a quantitative expression (though not always in accord with theory). All teams presented their data using a graph or table, half the groups using both Their attempts at quantification are generally not very successful from a scientific perspective, as they lack data analysis skills. This results in low ARPI scores in Table 6.5, that suggest little progress during the TLS. The information on progress, however, is less in what they accomplish and more in their effort:

> G4: If the height becomes twice as high, the fall time will be about 1.8 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.
> G11: From the results of this research we can conclude that the higher the cone starts, the longer it takes for the cone to reach the ground. There is also air resistance, but that doesn't really count in the results. The cone keeps the same mass, but each time travels a longer distance. The longer the distance, the longer it takes. Also, we can see from the results that if the cone falls from a twice as much distance, the time does not become twice as large.

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.

f. Critical attitude towards own approach and findings

Activity 1: We showed that in collecting their data, the students showed no critical stance towards their own approach, as they themselves admitted upon reflection. They were more critical in hindsight, when reporting to the stunt coordinator or considering doing the stunt 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 towards their own approach by, e.g., increasing the number of repeated measurements (all except G9 & 11), and deliberately choosing more accurate instruments (G1, G5, G9 & 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.

6.6 Discussion

Below we analyse to what extent integration of argumentation in inquiry succeeded, and resulted in development of selected UoE. We then establish whether the results show that students used what they had learned and satisfied the basic criteria to engage in QPI.

6.6.1 Evaluation of the implementation of design principles

DP1, the use of guided inquiry, was meant to ensure that students received enough guidance for maintaining progress in their inquiries and yet enough autonomy to use and evaluate their own ideas. We saw in all activities that students progressed smoothly and yet used their own approaches. They explored, guided by the teacher, the quality of the answers found and of the justifications they provided. We conclude that DP1 was implemented successfully.

DP2, minimizing the cognitive load, was meant to keep the activities so simple that distracting theoretical and procedural issues were avoided and enough time and energy remained to think and talk in class about how to obtain the best possible answer in the given circumstances. In activity 5 the interpretation of data patterns caused serious problems for students, but that was caused exactly by their own higher demands of the quality of the evidence. They showed a cognitive need for more PACKS type **B**&**C** knowledge that we see as a success. In no other activity did students experience cognitive overload, and yet the simplicity of the activities did not prevent students from developing non-trivial concepts of evidence such as fair test, variability, repeatability and outliers. We believe that the data show that DP2 was successfully implemented throughout.

DP3, the use of a real life context, was meant 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. In all activities we saw that reference to the context by the teacher helped students to attach meaning to the concepts and understandings of evidence. Students attached relevance to all contexts despite their being fictitious, and took them serious. That did not result in a spontaneous effort to obtain optimal scientific quality in their inquiry. But even a brief reminder of the episode created

in the 'Pirates' activity sufficed for students to reconsider the adequacy of their approach later on. Though not all intentions were realized, DP3 contributed in important ways to the integration of argumentation in inquiry.

DP4, the productive use of failure, is a principle that may easily be misunderstood. It is *not* about telling students what they did wrong, but about utilizing what *students* 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 that differed less from each other was in fact not a scientific failure at all, and used to discuss the CoE of natural variability in measurements. As our data show DP4 was successfully exploited in activities 1-4.

DP5, the use of meta-cognitive tasks, is indispensable in any teaching-learning activity that integrates argumentation. Arguments come into play only if claims are questioned, which requires reflection, the consideration of past statements and actions, contemplating their implications, 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 of 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 TLS. In the classroom the design principles actually interact and their combined implications are experienced.

6.6.2 Evaluation of the development of targeted UoE

UoE 6, the understanding that it is important to choose suitable instruments and procedures, was addressed in the TLS to ensure that students do not just use the first method that comes to their mind but consciously consider different methods of data-collection and procedures that would yield reliable and valid data. 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. We see no other viable explanation for their spontaneous request for help than an enhanced attainment of UoE 6, and a readiness for development of knowledge of type **B** and **C** based in that understanding.

UoE 8, the understanding that measured values will show inherent variation, should explain why it is unwise to take merely a single reading, or to follow a previously prescribed rule mindlessly. Students did repeat measurements in activity 1 but routinely, not with reason (Pols et al., 2022b). In activity 4 we confirmed that they relied on the 'naive' idea

that repeated measurements should yield the same result, an idea often reported in literature (Allie et al., 1998; Buffler, Allie, & Lubben, 2001; Lubben et al., 2001; Séré et al., 1993). In activity 5, their answers to the metacognitive task 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 that this number sufficed nor did they relate it to the variation in the measurements. However to develop more than an intuitive idea of what counts as 'enough' at this age is quite a tall order (Kok et al., 2019), as it requires a deeper understanding of how to quantify the variation and the ability to calculate and interpret measurement uncertainty (Lubben et al., 2001).

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. This shows inadequate attainment of UoE 9, the understanding that the range of values must be wide enough and the interval small enough to expose a pattern in the data. Students' choices of range and interval improved as all chose a range and interval that revealed a pattern. However, a justification for the choice remained absent.

UoE 14 is the understanding that a complete, clear, substantiated and useful answer to the research question must be formulated. Without it, the reason is missing for investing the time and energy to design an conduct a rigorous inquiry. Relatively high ARPI scores in activity 1 were mainly due to conclusions related to non-relationships that were easy to describe: students correctly stated 'that the results did not differ much'. Our data show that throughout the TLS students became aware that a conclusion must be as informative as possible. Drawing conclusions that were quantitative in nature, however, remained difficult due to a lack of knowledge in data analysis.

UoE 16 involves the understanding that the validity of conclusions does not go beyond the data available. It explains why a maximum range of the independent variable and size of the sample should be chosen. It clarifies the relevance of specifying the conditions under which conditions 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 an increased awareness of providing specific information on how the data were collected on which the claim is based. However, students did not improve their specifications of the limitations of the study.

6.6.3 Evaluation of progress in students' ability to engage in QPI

Unlike activity 1, the data of activity 5 indicate that students' inquiry, and their account thereof, can be regarded as an attempt to produce a scientifically convincing argument in support of a claim (criterion *a*). Though limited in extent and quality, students included *backings* in their letters.

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*). Those students who did not succeed in making scientifically more desirable decisions during the design of the inquiry were still able to specify such improvements in retrospect.

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

In the first activity students' approach can be described as 'controlled chaos' in which they seemingly unthinkingly gathered data to quickly 'get the job done'. Many teams progressed towards 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 TLS for obtaining 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 students searching for justifiable actions to their actually justifying them we consider to be one of the many next steps in teaching inquiry: enabling students to adequately justify and substantiate decisions pertaining to the targeted UoE, making all of their ideas about constructing evidence explicit to others.

6.7 Conclusions and future research

In this study we explored whether and how paying explicit attention to argumentation contributes to attain the highly valued learning goal of enabling students to engage in independent scientific inquiry. We showed that it is possible to have very young, inexperienced students to begin considering several core characteristics of scientific evidence that cause scientists to think and act in inquiry as they do, and develop an understanding of their relevance. We have shown that these students began to approach inquiry as the construction of a scientifically convincing argument by attempting to adhere, without being instructed to do so, to the specific UoE addressed in the TLS. The design principles that were used to develop the five activities that constitute the TLS result in students beginning to engage in argumentation in inquiry and develop an intention to produce the best possible answer in the given circumstances. Although operationalizing the

UoE in an integrated way in an independent inquiry remained difficult and their research did not improve much if judged by the traditional technical standards, students' words and actions showed that they were more sensitive to the quality of their research. 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 that.

We previously developed and validated a set of learning goals (the UoE) for integrating argumentation with inquiry (Pols et al., 2022a). We developed ARPI as an instrument for assessing the attainment of these learning goals. Here we extended and further clarified our recent framework for integrating argumentation in inquiry by linking the UoE and the *Procedural and Conceptual Knowledge in Science* model to Toulmin's argumentation model. We specified the UoE 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 on learning pathways towards these learning goals. We have shown that developing the UoE contributes to students' ability to engage in argumentation in inquiry, and that engaging in argumentation is promising to contribute to students' ability to engage in inquiry independently. We intend to further extend this theory and evaluate its value in enabling students to engage in scientific inquiry.

Further development of this proof of principle is certainly needed. In the first place, this study is based on a single TLS conducted in a single class taught by the teacher who developed the materials. While the amount and depth of the data needed in this paper hopefully clarify these limitations a test of the TLS in other settings is in order. Since teachers are reported to be ill equipped to teach scientific inquiry (Abrahams & Millar, 2008; Abrahams et al., 2014; Lunetta et al., 2007; T. J. M. Smits, 2003) and integrating argumentation puts further demands on teachers, the influence of the teacher needs further study, as does the development of ways to assist teachers. Familiarizing them with the framework for integrating argumentation in inquiry and providing exemplary activities could potentially help them. Another educational challenge is to find the time required to enable students to engage in inquiry. Collaborating with all science subjects would share this load and simultaneously illustrate how the inquiry insights are valued by all (natural) scientists, but research is needed how knowledge can effectively be transferred (Boohan, 2016a; Roorda, Vos, & Goedhart, 2015; Wong, 2017). Using our teacher education and in our in-service professionalization courses, we will further investigate these matters.

In our inquiry activities we reduced the number of decisions left open for the students and reduced the cognitive load pertaining to PACKS knowledge types **B** and **C**. These knowledge types inevitably interfere when students engage in more complex inquiries. Moreover, as inquiry is holistic in that each decision taken influences the next steps (Hodson, 2014), we face the challenge of enabling students to simultaneously apply several UoE in order to optimize their inquiry in terms of quality and time. Further research

is needed to explore how other UoE can be developed, in which order, at what level and to what extent. We have demonstrated that a better understanding of the nature of scientific evidence helps students to consider methodological choices but we need to explore how, e.g., the development of PACKS knowledge types **B** and **C** can be integrated. We think it is feasible to take next steps in achieving the highly valued but seemingly elusive goal of enabling secondary school students to engage in QPI on the basis of integrating argumentation in inquiry and look forward towards such progressions.

7. General conclusions, implications and recommendations

7.1 Problem statement and research questions

Despite many years of research we seem to hardly have made progress in improving the limited learning outcomes of practical work in secondary school science education. I started this study with the substantiated assumption that part of the limited learning outcomes of practical work in physics can be explained by the lack of students' ability to analyse experimental data (see chapter 1). The first study showed that students indeed exhibit an insufficient ability to analyse empirical data to engage in practical work independently. However, it also yielded a more worrying result: students are already hindered in an earlier stage which prevents them from successful engagement in practical work. They lack a sense of scientific purpose, which ought to be: to find and defend the best answer obtainable in the given circumstances. The outcome of this first study resulted in a shift of focus: from practical work in general towards *learning to engage in scientific inquiry*. It became a study towards the development of students' understandings of physics inquiry and fostering students' critical attitude in inquiry through a focus on argumentation. The idea was to build a teaching-learning sequence (TLS) that *progressively develops and refines students understanding of the purpose of scientific inquiries, and of the key concepts which underpin judgements about the quality of data* (Millar et al., 1994). The study explores the implications of paying explicit attention to argumentation in enabling students to engage in inquiry. The research questions guiding this study were:

> *1 What knowledge about scientific inquiry is required to plan, carry out and report a rigorous quantitative physics inquiry and how can mastery of this knowledge be assessed?*
>
> *2 What part of this knowledge has been thoroughly acquired by students who enter upper secondary school?*
>
> *3 What design principles are effective in guiding the design of a teaching-learning sequence that aims at enhancing students' critical attitude and developing inquiry knowledge through argumentation?*
>
> *4 What do students learn in a teaching-learning sequence directed at teaching inquiry through argumentation in terms of inquiry knowledge, enhanced critical attitude and use of argumentation?*

This chapter summarises the main outcomes of the four studies and presents the answers to the research questions above. The combined answers provide a substantiated answer to the main question:

> *What, and how, does paying attention to argumentation in inquiry contribute to enabling students to successfully engage in quantitative physics inquiry?*

The chapter is concluded by specifying possible pathways for future studies and recommendations for education.

7.2 Quantitative physics inquiry – its aims and their assessment

One of the main learning goals of secondary school physics education is *learning to engage in scientific inquiry*. In this study we narrowed this broad learning goal and focussed only on inquiries in which the relationship between two physical quantities is to be determined. This type of inquiry is referred to as quantitative physics inquiry (QPI). Although one might have an idea of what students should know and be able to do when planning, conducting and reporting a QPI, questions are what it *precisely* entails and how mastery of this knowledge can be assessed. This led to the research question:

> *What knowledge about scientific inquiry is required to plan, carry out and report a rigorous quantitative physics inquiry and how can mastery of this knowledge be assessed?*

In their *Procedural and Conceptual Knowledge in Science* (PACKS) model, Millar et al. (1994) distinguish four types of knowledge required to successfully engage in QPI. Knowledge type **A** pertains to the understanding of the nature and purpose of the task, knowledge type **B** to the relevant conceptual knowledge, knowledge type **C** to the understanding of measurement instruments and students' ability to use these, and knowledge type **D** to the understanding of criteria for evaluating the quality of empirical evidence. According to Millar et al. (1994), knowledge type **D** crucially influences the quality of the research. Although all types of knowledge need to be developed in order to enable students to plan, carry out and report a rigorous QPI, our attention focused on acquiring knowledge type **D**. In chapter 3 we specified the Understandings of Evidence (UoE) – insights and views that an experimental researcher relies on in constructing and evaluating scientific evidence – which enhance knowledge type **D**. We regard these UoE as important learning goals for inquiry teaching in physics education as these understandings are used in achieving and assessing the scientific cogency of the argument in support of a claim. The UoE express how knowledge of scientific 'rules' in the various phases of an inquiry are applied in observable actions and decisions, such as drawing a conclusion, and how these help in constructing a scientifically cogent argument. Our set of UoE (and its framework) comprehends thus an important part of the answer to the first part of the research question. We should augment here that the UoE can further be divided in other (sub)competences or insights. For

instance, what UoE 12 (*data require appropriate methods for analysing and describing them*) entails at lower secondary school level was specified in the first study (chapter 2).

Mastery of inquiry knowledge, specifically students' UoE, can be tested and assessed in various ways. We have shown that engaging students in more authentic QPI reveals much of their inquiry knowledge. Assigning attainment levels of inquiry knowledge (the UoE) is possible because the UoE have been precisely defined and indicators for various attainment levels have been developed and validated. The minimum attainment level of an UoE can be assigned under the condition that the assessor has an appropriate attainment level; enough time for assessment is available; and the relevant information is available in the form of, e.g., an oral discussion, a lab journal or a report. Acquiring the relevant information has proven to be difficult when young, inexperienced students engage in a more authentic QPI: their substantiation of decisions – in either the scientific graphic organizer (SGO) or report – is limited in quality and extent. Their thinking behind the doing remains often implicit. This means that if we want to assess students' understanding with more precision and certainty, we need either to develop students' awareness that choices have to be explicitly substantiated, or look for other ways of assessment rather than school science reports. For instance, the tasks that instigated discussion and the meta-cognitive tasks in the activities revealed students' understanding of the targeted UoE as well.

But even with a 'list' of what to assess, and indicators for specific levels, assessment of mastery of a single UoE remains difficult. What the best decision is, is a matter of judgement (Lipton, 2003). It requires interpretation of the assessor, for instance: Is the number of repeated measurements adequate in light of the *practicality of consequences* and the *spread in measurements*? Moreover, in more complex inquiries, interference between UoE, and between UoE and physics content knowledge, is inevitable. Although this could be perceived as a downside of the UoE and the assessment method, we believe that it can and should be perceived as a plus. Rather than degenerating inquiry at secondary school level into a set of specific tasks described in a checklist (with students memorizing each item) – where the highest level is ticked when, e.g., students repeat a measurement five times – students still have to think whether measurements need to be repeated and if so, how many repeated measurements are sufficient. This assessment format resembles the process of scientific inquiry and peer review: There is no single unique scientific method that always works.

7.3 Quantitative physics inquiry – Students' understandings of and ability to engage in

Using the UoE as our focus of enabling students to engage in QPI, we can answer the question what students already know of doing QPI:

> *What part of this knowledge has been thoroughly acquired by students who enter upper secondary school?*

Studies 1 and 3 indicated that students have a rudimentary grasp of inquiry knowledge. They know that only a single variable is to be changed, a measurement is to be repeated, that the average value is to be calculated and that their measurements should result in a conclusion that answers the research question. This knowledge and associated actions stem mainly from the teacher telling them to do so rather than that these are based on a true understanding of the concepts at hand. That is, they understand *that* they have to repeat a measurement but do not specifically understand *why* they should do so. Moreover, students at the age of 14-16 still lack the understanding why the experiment should result in *the most informative conclusion that is supported by the data* and that they have to show that their answer is the best available. Studies 1 and 3 illustrated that students were quite satisfied with a common sense or superficial answer to the research question. However, studies 3 and 4 illustrate that we can teach students not only that certain 'rules and procedures' need to be adhered to, but also *why* they should do so, what purpose is being served in doing so. Knowing why, in addition to knowing that, has brought students into the position in which they want to use knowledge that they did not acquire yet: Students attempted to produce a conclusion that is quantitative in nature and that is backed by their data in the final inquiry but they still lacked the ability to properly analyze their data to adequately determine and describe the quantitative relationship between two quantities. Offering that knowledge in subsequent activities has a chance of being successfully processed as students themselves have a cognitive demand to develop that knowledge.

From these findings we learn that if we want students to be able to independently plan and conduct a basic QPI we have to enhance students' understanding of each of the UoE *and* of the purpose of scientific physics inquiry and develop the associated critical attitude, as was attempted with our TLS.

7.4 What design principles are effective in guiding the design of the TLS

Progress in education is likely to be made if we can develop activities that are guided by design principles that have proven to be effective in instigating desired actions in students. Moreover, as stated by P. Bell et al. (2004): *In education, it is becoming increasingly common to represent design knowledge and theoretical insights as design principles that emerge from research with the goal of informing future design activities.* Therefore we asked the question:

> *What design principles are effective in guiding the design of a teaching-learning sequence that aims at enhancing students' critical attitude and developing inquiry knowledge through argumentation?*

The extensive literature on teaching scientific inquiry and practical work helped in identifying and formulating design principles for the stated aims. Especially the *Procedural and Conceptual Knowledge in Science* (PACKS) model played a central role in guiding our design decisions. As the design principles have been extensively described throughout the chapters, we merely summarize these along with the established merits and limitations:

Guided inquiry Learning to engage in inquiry requires that students make decisions, try, fail, reflect and try again. In guided inquiry students carry out the inquiry as they seem fit and help is offered when asked for, though the decisions remain theirs to make. Students can learn much from the struggles they encounter, provided they understand and accept the purpose and intent of scientific inquiry. However, studies 1 and 3 show that students hardly ask help, do not consider the adequacy of their choices, do not critically reflect on their approach and they certainly will not ever try all over again. The central reason in our analysis is: they have no reason to be concerned about the quality of their answer, and no way to assess that quality. Only once students want to produce a quality answer, they start to consciously recognize what choices need to be made and consider what are scientifically adequate decisions. They then turn to the teacher for assistance if there is no readily available solution. Students thus require, first, a critical attitude which spurs them to critically reflect on their own decisions before the possibilities offered in guided inquiry are optimally exploited.

Reduction of cognitive load other than PACKS knowledge type D Students are often put in situations where their inadequate skill level acts as a considerable barrier to learning (Hodson, 1990). For young, inexperienced students, these barriers should be lowered, e.g., by minimizing distracting factors. Using the PACKS model and our goal to develop PACKS knowledge type **D** specifically, this means that students clearly understand the task, are familiar with the equipment or are helped to become familiar with it, and students master

the content or the content is not relevant. Our studies show that, by reducing the cognitive load, students can carry out the given task and still encounter many difficulties. However, these are foremost the difficulties we wanted them to encounter. We should note here that to further enable students to engage in QPI it is likely that students ought to engage in more complex QPI, where they apply in an integrated way the full scope of PACKS.

Contextualized problems Contextualizing the activities, we reasoned, could elucidate the relevance of producing quality work. This would potentially enhance students' critical attitude towards their own work. However the three intervention studies showed that merely contextualizing the inquiry hardly enhanced students' intrinsic critical attitude. Still the context helped the teacher in making abstract concepts tangible. Moreover, it helped in showing the relevance of producing reliable data and useful conclusions. The context, as used by the teacher, thus served as a tool to, extrinsically, enhance students' critical attitude.

Producers and consumers of knowledge The context was optimally exploited with the design principle *produces and consumers of knowledge*. By switching students' perspective to that of a consumer who is potentially exposed to the severe risks of applying the findings in the real world, students critically engaged in reviewing the adequacy of their own inquiries. As described in study 3, this significantly changes students' view on scientific inquiry. Applying this design principle fostered students' critical attitude towards their own approach to inquiry, and incited them to acquire inquiry knowledge in the subsequent activities. The design principle was not as strongly applied in the subsequent activities, but as an episode was created in the first activity of the TLS, reminding students of the potential consequences implied by the context sufficed to have them consider the adequacy or implications of their decisions.

Productive failure In scientific research, researchers understand that a scientifically convincing argument for the future claim should be produced. They look for information on questions that they cannot directly or convincingly answer, and as a consequence learn. And if not, the reviewer will point out shortcomings and they will learn from their 'failures'. However, in secondary school science both the understanding that and why the best answer should be produced is lacking, as well as the relevance (and therefore the motivation) to produce an informative conclusion. These issues create barriers to self-directed learning. Studies 3 and 4, however, show that making productive use of students' 'failures' is an effective way to address important concepts in inquiry. In addressing the weaknesses in students' approaches, these become the center of attention and learning takes place through discussion with and explanation from the teacher.

Meta-cognitive tasks The value of meta-cognitive tasks in learning seems undisputed (Livingston, 2003). However, in practical work such activities are rarely utilized. Students

formulating 'rules for doing proper inquiry' and reflecting on how these 'rules' could have improved their inquiry in the first activity has shown to be a fruitful way to engage students in such meta-cognitive tasks. It not only consolidates learning and shows how the content is relevant to the students, these rules also provide some guidance in setting up a new inquiry. Moreover, students' answers provided valuable information for the teacher (formative assessment).

7.5 What students learn

The TLS was developed to enhance students' understanding of the scientific purpose of inquiry and specific UoE, and enhance their critical attitude – crucial in considering what are scientifically adequate decisions. Although we did not expect students to become proficient researchers, to what extent we could yield the desired outcomes when focussing on argumentation was unknown. The relevant question was therefore:

> *What do students learn in a teaching-learning sequence directed at teaching inquiry through argumentation in terms of inquiry knowledge, enhanced critical attitude and use of argumentation?*

The inquiry carried out in the final activity of the TLS and students' responses to various questions in the three instructional activities, show that students have a firmer grasp of what conducting a QPI entails and that they developed various UoE – albeit at a rudimentary level. Important progress in enabling them to engage in inquiry is their enhanced awareness that (in inquiry) doing without thinking is pointless. In other words, they developed the understanding that answers to research questions are only useful if these are trustworthy and optimally informative. This enhanced their critical attitude towards their own approach. That is to say, in the first inquiry a critical attitude appeared to be almost completely lacking but at the end of the TLS students recognized what decisions were to be made and deliberately attempted to make adequate ones. Students seemed to be more sensitive to the quality of their research and tried to produce a reliable, useful and informative answer to the research question. Moreover, from both their approach to the given problem and their reports to NASA it follows that most students made progress in constructing the inquiry as a scientifically cogent argument in support of a claim. In their reports, several teams elaborated on what was done, how it was done and provided a substantiation for their decisions, though brief and often incomplete – as can be expected of students of this age.

7.6 Developing inquiry knowledge through argumentation

In chapter 1 we made a plea for a focus on argumentation when teaching students how to engage in QPI. Our implicit working hypothesis was that students need to understand why only the best available answer in the given circumstances is truly satisfactory and understand that they ought to convince themselves (and others) that they did produce that answer before we can expect them to use argumentation to improve and defend their inquiry. We are now in the position to reflect on the broader research question:

> *What, and how, does paying attention to argumentation in inquiry contribute to enabling students to successfully engage in quantitative physics inquiry?*

We will answer this question using different perspectives: from a philosophical point of view, from the students' perspective, and from the teacher's perspective.

The philosophical perspective In each chapter of this thesis we elaborated on the central role of argumentation in inquiry: In setting up the inquiry, the researcher uses argumentation to convince others why answering the posed question is relevant, what knowledge gaps are to be filled and why that is worth the effort. In conducting the experiments the researcher uses argumentation to evaluate and improve the quality of the study. The inquiry can be considered to be finished when the peers have scrutinized it, when they have considered whether the claim is substantiated well enough so that it is accepted. In presenting their verdict, these experts provide arguments that support their decision. In each in-between step of the inquiry, the researcher has to decide and evaluate what has to be done to support of the claim, making the claim as indisputable as possible.

Therefore, if we neglect argumentation in teaching inquiry, we would provide a 'false' image of how science works. And indeed, various studies suggest that closed, prescribed inquiries are counterproductive in developing students' understanding and appreciation of scientific methodologies (Ansell & Selen, 2016; Wilcox & Lewandowski, 2016; Zwickl et al., 2014). Reversing this issue provides the answer to the *what* question from a philosophical point of view: focusing on argumentation in inquiry provides a more authentic and realistic image of doing science. Argumentation is the start and the end of a scientific inquiry. Therefore, a focus on argumentation in teaching inquiry is indispensable.

The students' perspective The study showed that integrating argumentation in inquiry through an emphasis on evidence enhanced students' awareness of the purpose of scientific inquiry: that they ought to produce the *best available answer* to the research question. In turn, this enhanced understanding fostered students' critical attitude towards their inquiry approach. Through the focus on the quality of the evidence in the instructional activities, students gradually developed basic inquiry insights. Ultimately, though in a

limited and implicit way, students started considering the question "what is the best next step?".

From the students' perspective, the answer to the *what* question is that attention to argumentation contributes to the development of a critical attitude towards students' own approach and fosters the development of inquiry knowledge. The answer to the *how* question is that attention for argumentation implies a continuous focus on the quality of the evidence and the claim. It requires students to evaluate whether what they think, say, do or claim is correct or (scientifically) justifiable. Through discussions and teacher's explanations they developed a better understanding of the quality of evidence.

The teacher's perspective A neglected aspect in this study is the teacher's (or the designer's) perspective. The following text is a more personal reflection that is relevant for teaching inquiry. In designing and teaching each of the activities, the idea of producing a scientifically cogent answer has been helpful, for instance in considering and evaluating different contexts and approaches. We considered, e.g., the context of designing a game for activity 4 of the TLS. However, the context of car safety would put a stronger emphasis on the quality of the data and was therefore preferred. The choice for this particular context turned out well: it helped students to evaluate the potential consequence of inconsiderately discarding outliers.

The answers to the *what* and *how* questions from the teacher's perspective are thus that a focus on argumentation contributes to the development of meaningful activities in which students develop inquiry knowledge. Moreover, attention to argumentation contributes to guiding the students during the activities because the focus on the quality of evidence and the tenability of the claim can be used to have students critically evaluate their own work. Students do not (or hardly) need the teacher's verdict when they are asked whether the findings are reliable or the conclusions useful. They are often able to judge the quality of the inquiry themselves when their attention is turned towards evaluating the quality of their work. A central evaluation in which the inquiry's quality was discussed resulted in a rich discussion in which various, abstract, concepts came forward and became more tangible to students. In subsequent practical activities that I carried out in class – not described in this thesis – I experienced that the questions 'are you convinced of …' or 'can you convince me that …' sufficed to have students evaluate their decisions and reveal their understanding.

7.7 Schematic of the theoretical framework

In the chapter 1 we introduced the Toulmin (2003) argumentation model, the PACKS model and the Concepts of Evidence. We surmised that these different models and concepts are somehow connected, but could not state how they did so. We are now in a position in which

7.7 Schematic of the theoretical framework

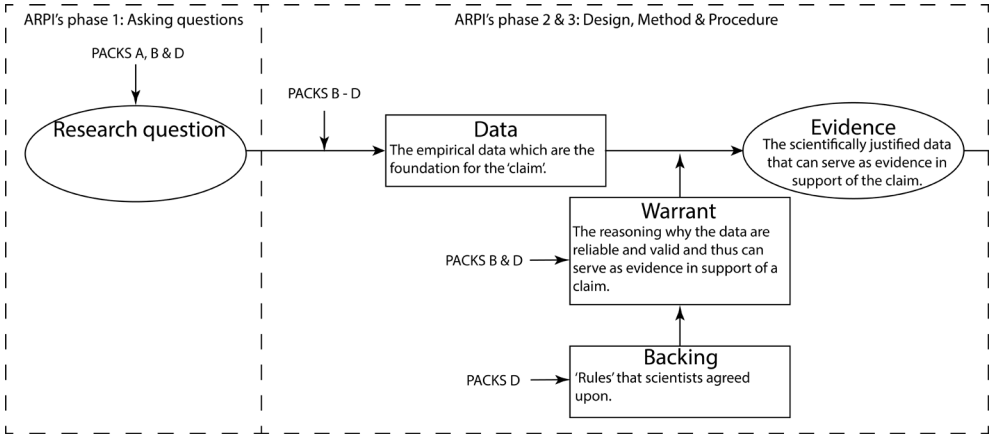


Figure 7.1 A schematic overview of the coherence between the various theoretical models and

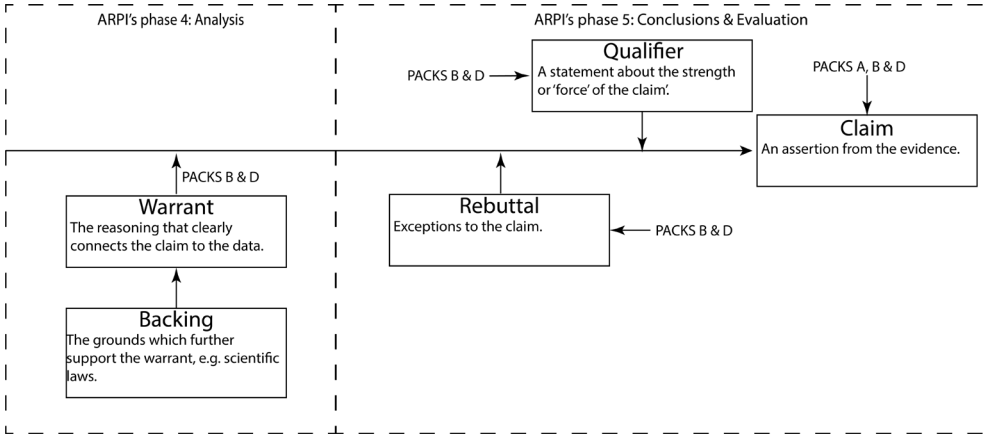
we can look back and try to connect these. That is, provide a schematic and coherent overview of the theoretical framework.

In figure 7.1 elements of Millar's (1994) PACKS model and the Toulmin (2003) argumentation model are recognizable. The rectangles contain the field-invariant elements from the Toulmin argumentation model. The ellipses stem from the PACKS model and have been added to make the Toulmin argumentation model fit well with experimental physics research. The dotted squares indicate the phases of ARPI, where each of these phases include various UoE (see chapter 3). The various CoE are recognizable in the UoE.

The dashed lines seem to show the borders of an inquiry phase. However, these borders are more or less arbitrary. Moreover, as in the PACKS model, various vectors indicating back-loops have been omitted: doing a scientific inquiry is not a linear process starting from a research question straight towards a conclusion. The schematic overview is thus a model: a simplification of the complex process called inquiry. Still, the model provides an accurate picture of how an argument is often constructed and reported in a scientific paper.

An inquiry starts with a research question (ARPI's phase 1: Asking questions) (J. S. Lederman et al., 2014). The interpretation of the specific question is influenced by PACKS knowledge type **A**, **B** & **D** (Millar et al., 1994): the understanding of the nature of the task, the relevant conceptual knowledge that further frames the question as well as the understanding of the quality of evidence. We have identified three understandings of evidence (UoE 1-3) that are relevant in this phase.

Based on the research question, an experiment is designed which provides an answer to the research question (ARPI's phase 2: Design). Data are collected by conducting the experiment (ARPI's phase 3: Method & Procedure). In both designing the experiment and conducting it, PACKS knowledge type **B**-**D** are prevalent. Those data only become



elements used and developed in this study.

evidence once the quality of the data is interpreted, after a justification of the design, methods and procedures. This justification thus requires a warrant which can be further supported by backings: the 'scientific' rules for doing inquiry that scientists in that specific field agree upon. In providing warrants, PACKS knowledge type **B** & **D** are used. Moreover, in designing an experiment, devising a method and procedures and justifying choices UOE 4-10 are relevant.

These data are further analysed (ARPI's phase 4: Analysis UoE11-13), resulting in a claim: an answer to the research question (ARPI's phase 5: Conclusions & Evaluations). In stating qualifiers and rebuttals, PACKS knowledge types **B** & **D** are again prevalent: the conditions under which the claim is true (at least those specified by the researcher) are determined by the knowledge of the physical concepts and the knowledge of the quality of the data. Knowledge of the quality of the data, the methods used, but as well as the limitations to our theoretical knowledge inform these Toulmin field-invariant elements. Moreover, in this process of drawing conclusions and setting limitations to the validity of the claim, UoE 14-18 are the relevant insights upon which a researcher relies.

This summary of the combination of the two models and the theoretical concepts developed in this study surely needs further scrutiny. Still, it might provide a point of departure for a next study into the integration of argumentation in inquiry.

7.8 Directions for future research

This study shows, to a large extent, what we are able to achieve with students at the end of lower secondary education in terms of developing inquiry knowledge within the constraints that come within secondary school physics. Many aspects of teaching secondary school students how to plan, carry out and report a rigorous QPI were addressed and investigated.

The aspects that were not addressed – the limitations of this study – and the insights that were gained provide directions for future research.

At the end of chapter 3 we coined the idea of a pedagogical theory for teaching scientific inquiry on the basis of argumentation. This theory is built on the framework of argumentation in inquiry, with the UoE being the specified learning goals. In chapter 5 we extended this theory by (more clearly) linking the UoE and the *Procedural and Conceptual Knowledge in Science* model to Toulmin's argumentation model, formulating design principles that integrate argumentation in inquiry in an educational setting, and developing and testing an educational approach for teaching scientific inquiry. In short, for many aspects of teaching scientific inquiry, this study provides basic elements that require further practical and theoretical development. In this sense, the pedagogical-didactic basis for a hitherto missing part of 'learning to engage in inquiry' has been laid. This pedagogical-didactic basis needs to be further extended, optimized and tested. For instance, what are the limitations to ARPI with regard to deriving students' understanding from their actions and justifications? And how can we (re)design practical activities so that the focus is on justifying decisions rather than complying to rules? And if we succeed in redesigning these activities, what and how do these activities contribute to learning?

We identified the understandings required to plan, conduct and evaluate a scientifically rigorous QPI. We designed an assessment format that allows us to determine students' attainment level regarding these understandings. Subsequently we developed and tested the TLS where *some* of these understandings were the targeted learning goals. But as stated, to produce a meaningful answer one has to pay attention to detail across *all* of ARPI's categories (Pols et al., 2022a). Not addressing all understandings limits our study. The TLS can (and should) be improved and extended. However, carrying out an extended TLS would consume considerable time of secondary physics education. This issue could be overcome when teaching inquiry is not limited to the subject of physics and teachers from various science subjects collaboratively work on attaining this learning goal. This is possible as the UoE are potentially relevant to all natural sciences. It would probably help education a great deal if the science subjects collaboratively develop a deeper understanding of scientific inquiry in students and together teach them how to plan, conduct and evaluate a scientific inquiry. Note that here 'physics' is replaced by 'scientific'. A broad question worthy to investigate is:

> *How can the science subjects collaborate to develop a deeper understanding of scientific inquiry in students?*

As argued in chapter 1, more access to pertinent knowledge leads to higher quality inquiries. However, this study did not include an investigation on whether the TLS affects students' approach to practical work (in general) in the long term. Whether students' enhanced understanding of doing inquiry indeed improves the learning outcomes of

practical work in general, is yet unknown. There is anecdotal evidence that students who engaged in the TLS carried out subsequent practical activities more purposefully. These students seemed to have acquired a better understanding of what is expected of them and, importantly, seemed to think before doing. Whether students' enhanced understandings indeed result in better learning within other practical work, whether further incentives are required to have students maintain scientific standards in non-contextualized practical work, is to be investigated. A question worthy to investigate is therefore:

> *How does a teaching-learning sequence with a focus on developing students' understanding of inquiry impact the learning outcomes of conceptual practical activities?*

This study is limited by the choice for case studies, where each of the activities was carried out by the same teacher. Although we did not extensively report it, other teachers have carried out the TLS as well. We summarize here that all of them were content with the TLS, and most of them had similar findings as described in chapters 5 and 6. They all expressed the intent to implement the TLS in future years. This may indicate that the materials are sufficiently accessible for other teachers. However, they might have had good experiences because they already had interest in – and experience with – teaching QPI. Whether similar learning outcomes may be expected from less experienced teachers remains a question, especially as it is reported that many teachers are not well equipped to give substance to the learning goal *learning to engage in scientific inquiry* (Abrahams & Millar, 2008; Abrahams et al., 2014; Crawford, 2014; Lunetta et al., 2007; T. J. M. Smits, 2003; Spaan, Oostdam, Schuitema, & Pijls, 2022). This, augmented by the studies that report that the teacher plays an important, if not crucial, role during practical work (Abrahams & Millar, 2008; Dillon, 2008; Tamir, 1991; Van Rens & Dekkers, 2000; Watson, Swain, & McRobbie, 1999), provides reason to further study what knowledge and materials teachers require to teach scientific inquiry. There seems to be a need to further develop professional development courses for teachers and investigate the effect of these activities in order to optimize such courses. A question worthy to investigate is therefore:

> *How do we optimally enable teachers to give substance to the learning goal "learning to engage in scientific inquiry"?*

I think it is worthy to mention that the associated teaching materials developed in this study are used by various science teacher educators in the Netherlands. There is even evidence that the translated materials (Pols, 2021a) are used in national professional development courses in France (Hihi, 2020). This illustrates that the material is believed to have potential to educate pre-service teachers. Further investigating the outcomes of such interventions is worthwhile.

This study focused on enabling students to plan, carry out and report a rigorous QPI. But evaluating the quality of the work of others is an important aspect of doing science as well. We did not yet focus on students' ability to determine the quality of the work of others, students were foremost asked to evaluate the quality of their *own* inquiries. As peer review is a powerful tool to maintain and enhance the scientific quality of inquiries, a question worthy to investigate is:

> *How can peer review be used to improve the quality of the inquiry and help students in understanding and adhering to the scientific standards?*

Although I did not present it, we made a start with investigating how peer review can be utilized to have students uphold scientific standards. Rather than confronting students with the consequences of their own 'flawed' inquiries, they were asked to judge the quality of the work of others. In the role of employers that commissioned the investigation, the students were asked whether they accept the findings and claims of their fellow students who actually conducted the inquiry. Peer review improved students' approach to subsequent inquiries. The article presenting this idea and the findings will be, hopefully, soon at your disposal.

In my current job as coordinator of the first year physics lab course, I have 300 students and their inquiries at my disposal for future research. Currently, I implement the gained insights in redesigning the course and simultaneously 'record' the effects of operationalizing these ideas. I hope this study and future studies will improve physics education with regard to teaching scientific inquiry. I conclude this chapter with recommendations and implications for education.

7.9 Recommendations and implications for education

The first educational recommendation is that if we really value the important outcomes that *learning to engage in scientific inquiry* in secondary science education has or might have, then we should make more effort in enabling students to do so. If only for the sake of scientific literacy – that is the general scientific awareness and understanding of science that the general public needs to have (Durant, 1994; Osborne, 2000). The general fuss about the COVID-19 measures illustrates that there is still much to be gained with regard to scientific literacy. If everyone should have a basic knowledge about science, (lower) secondary education seems to be the place to attain this goal (Duschl, 2008; Gott & Duggan, 2003; Hurd, 1998; Laugksch, 2000; Millar, 2004; Millar & Osborne, 1998; Osborne, 2000; D. A. Roberts & Bybee, 2014). Engaging students in scientific inquiry is one, but important, way to raise students' understanding of the scientific processes (Gott & Duggan, 2007; Hofstein & Kind, 2012; Kanari & Millar, 2004; Millar et al., 1999). Studies 1 & 3 indicate that students are still far from becoming scientifically literate persons. Studies 3 & 4 illustrate that we are

capable in raising students' awareness of the quality of the outcomes of scientific investigations and have them value it. These studies also demonstrate that we can engage students in scientific inquiry. However, enabling students to engage in scientific inquiry takes a considerable amount of time. This is an issue as time available is one of the two things that is frequently lacking in secondary school education (the other being the financial means).

The second recommendation is the suggestion to reconsider and rephrase the learning goals related to *learning to engage in scientific inquiry* in curriculum documents. If we compare national and international curriculum documents and follow-up on the recommendations of various scholars, it follows that at the end of secondary physics education students should be able to devise and conduct a basic physics inquiry with a large degree of independence (HMC Eijkelhof & Kortland, 2001; Hodson, 1990, 1993, 2001, 2014; Hofstein & Kind, 2012; Holmes & Wieman, 2018; Lunetta et al., 2007; Millar, 2008; Netherlands Institute for Curriculum Development, 2016; Ottevanger et al., 2014; Singapore, 2019; United Kingdom Department for Education, 2014). These requirements are in line with the idea of raising scientific literacy. What devising and conducting a basic physics inquiry encompasses, seems clearly described by the Netherlands Institute for Curriculum Development (2016) (author's translation):

> *A student should be able to use consistent reasoning and relevant mathematical skills in order to answer the research question. The student should be able to make observations and collect relevant data; process and present the data in such a way that it helps to answer the research question; draw conclusions which are based on the data; evaluate the execution and conclusion using the terms validity, accuracy, reproducibility and reliability.*

However, these learning goals are formulated in such way that the attainment level depends on the user's interpretation. If we take these requirements serious, this would imply that students are professional researchers once they finish secondary education. The abilities described are in essence the abilities of an experimental physicist. However, without any specifications of the attainment levels, one could also advocate that these (learning) goals are (too) easily attained: students of all ages are able to draw conclusions based on data. One should take it for granted that such conclusions are not informative (as demonstrated in chapters 2 and 5). Choosing either one of the extremes (becoming an expert or staying an apprentice) can hardly be the intention of the curriculum developers. We cannot expect students to have fully attained all of these abilities at an expert level. But as well, we cannot be satisfied with students drawing only qualitative, superficial conclusions. There thus needs to be better specified what we expect of our students. This could be done by providing examples of experiments that students could conduct independently, as was done by Mooldijk et al. (2006). Or by further explicating what *kind* of

conclusions are sought. For instance, the first sentence of the Dutch physics curriculum pertaining to scientific inquiry could be rephrased on the basis of UoE 14: *A students should be able to produce a scientifically convincing argument for a conclusion that is optimally informative*. Our UoE and its framework might be useful in specifying what then is required to do so, and subsequently formulating (rephrasing) learning goals.

The final recommendation is that teacher educators are to use the PACKS-model to teach prospective science teachers about the pedagogy of practical work. The model clearly distinguishes the three (main) learning goals for practical work often mentioned in literature (Dillon, 2008; Hodson, 1990; Millar, 1998, 2004; Tamir, 1991; Wellington, 2002):

1 Developing knowledge and understanding of science: learning about specific science concepts. (PACKS knowledge type **B**)
2 Developing practical skills: learning to handle and manipulate equipment. (PACKS knowledge type **C**)
3 Developing scientific inquiry and process skills: learning how to do scientific research. (PACKS knowledge type **D**)

The model has helped me to specify learning goals, reduce the cognitive load in the activities, guide scaffolding comments while teaching and so on. It has even enabled me to renew an entire physics lab course at the Delft University of Technology. If we want to improve the learning outcomes of practical work, we have to equip prospective teachers with a firmer theoretical foundation of the pedagogy of practical work. If not, progression will be slow as teachers tend to teach in the same way as they were taught.

8. Summary

8.1 Samenvatting in het Nederlands

Leren onderzoeken is een belangrijk leerdoel van natuurkundeonderwijs op de middelbare school. Vrijwel gelijk aan andere (internationale) curricula staat in het Nederlands natuurkunde curriculum dat leerlingen een onderzoek moeten kunnen uitvoeren en conclusies kunnen trekken waarbij consistente redeneringen gebruikt worden. Echter, de literatuur duidt dat dit een schijnbaar onhaalbaar leerdoel is. Dit promotieonderzoek richt zich dan ook op de vraag hoe we in leerlingen de kennis die nodig is om een 'gedegen' kwantitatief natuurkundeonderzoek (KNO) – het veel gebruikte type onderzoek in middelbare school natuurkunde waarin het verband tussen twee grootheden wordt bepaald – op te zetten, uit te voeren en te evalueren, effectief kunnen ontwikkelen. KNO dekt veel, maar niet alle soorten onderzoek binnen natuurkunde en *leren onderzoeken* behelst veel meer dan het precies vast stellen van de relatie tussen twee grootheden – bijvoorbeeld ook het kunnen koppelen van die relatie aan bestaande theorie. Echter, als het lukt om leerlingen een gedegen KNO op te laten zetten, kunnen we van daaruit andere belangrijke elementen van *leren onderzoeken* afdekken. Belangrijk hierin is het begrijpen waarom de gekozen aanpak (in tegenstelling tot eerdere pogingen) werkt, en op die manier onze theoretische kennis over het onderwijzen van *leren onderzoeken* uit te breiden.

De eerste studie richt zich op de vraag wat leerlingen al weten over het doen van KNO, specifiek met betrekking tot het analyseren van empirische data en het trekken van optimaal informatieve conclusies. Op basis van curricula documenten specificeren we wat leerlingen aan het eind van de onderbouw moeten kunnen. Middels practica en gerelateerde activiteiten onderzoeken we of ze dat ook kunnen. Uit het onderzoek trekken we de conclusie dat het kennisniveau van de onderzochte 4-HAVO/VWO leerlingen niet op het te verwachten niveau is. Ook leiden we uit dit onderzoek af dat we leerlingen moeten motiveren om een volledig, juist en onderbouwd antwoord op de onderzoeksvraag te produceren. Pas als ze dat zelf echt willen, zullen ze het probleem van het beperkte kennisniveau erkennen en daar iets aan willen doen. Dit onderzoek vult het PACKS model (*Procedural And Conceptual Knowledge in Science)* – waarin vier typen kennis onderscheden worden die nodig zijn om QPI uit te voeren – aan door de noodzaak van integratie van argumentatie in leren onderzoeken te duiden. De rol van argumentatie in *leren onderzoeken* is essentieel maar bleef tot nu toe onderbelicht in onderwijs.

In de tweede studie beschouwen we onderzoeken als het construeren van een overtuigend argument, en identificeren we de kennis waarop een onderzoeker vertrouwt bij het construeren en beoordelen van de overtuigingskracht wanneer het gaat om KNO. Die kennis, de zogenaamde *Understandings of Evidence* (UoE), beschouwen we als belangrijke leerdoelen voor *leren onderzoeken*. Omdat leerdoelen slechts waarde krijgen

als we in staat zijn om het verworven niveau van leerlingen vast te stellen, specificeren we per UoE verschillende niveaus met bijbehorende acties en beslissingen waaruit dat niveau af te leiden valt. Op het hoogste niveau gebruikt de leerling argumentatie om de gemaakte keuze of actie te rechtvaardigen. Het zo ontstane construct bestaande uit leerdoelen en indicatoren voor verschillende niveaus, de Assessment Rubric for Physics Inquiry (ARPI), is in deze Delphi studie gevalideerd. Het onderzoek biedt een kader voor integratie van argumentatie en onderzoek.

Op basis van dat kader en de verworven inzichten uit de eerste twee studies ontwikkelen we de lessenserie die beschreven staat in hoofdstuk vier. Deze lessenserie is onderwerp van onderzoek in de laatste twee studies.

Om leerlingen te overtuigen om de benodigde inspanning te leveren die nodig is om een wetenschappelijk adequaat antwoord te produceren en in hen de benodigde onderzoekskennis te ontwikkelen, moeten we eerst bij hen de behoefte creëren om zo'n antwoord te produceren. In deze derde studie, die zich richt op de eerste activiteit van de lessenserie, proberen we dat te doen door de uit te voeren KNO te contextualiseren en leerlingen de kwaliteit van hun onderzoek te laten evalueren vanuit de gebruiker die de eventuele consequenties van slecht onderzoek ondergaat. Dit levert op dat leerlingen op basis van hun eigen criteria de kwaliteit van hun onderzoek als onvoldoende beschouwen. De leerlingen zien in dat wetenschappelijke standaarden nodig zijn wil een conclusie echt bruikbaar zijn. Wat die wetenschappelijke standaarden inhouden, is hen op dat moment nog wel onbekend, maar ze lijken bereid die kennis te willen ontwikkelen.

Die bereidheid wordt gebruikt om in de volgende drie activiteiten kennis over onderzoeken te ontwikkelen. Om argumentatie in die activiteiten te integreren en de kritische houding van de leerlingen verder te ontwikkelen, hebben we vijf ontwerpprincipes geformuleerd die we in elk van de activiteiten implementeren. De effectiviteit van die principes is bepaald door na te gaan of de verwachte opbrengst in elke activiteit gerealiseerd wordt. De laatste activiteit, een KNO, wordt gebruikt om na te gaan wat de lessenserie en een focus op argumentatie oplevert in de zin van ontwikkelde onderzoekskennis, kritische houding en gebruik van argumentatie.

Door systematisch de zwaktes in de aanpak van de leerlingen (design principe 4) in de gecontextualiseerde, eenvoudige onderzoeksactiviteiten (design principe 1-3) aan te pakken en hen te laten evalueren wat die kennis is en hoe die gebruikt had kunnen worden in hun eerste KNO (design principe 5) vindt leren plaats. Het gevolg is dat leerlingen in de laatste KNO nadenken over hoe ze bepaalde onderzoekskeuzes (bijv. de keuze voor een meetinstrument) het best in kunnen vullen. De leerlingen doen een goede poging om een informatieve conclusies te produceren en te onderbouwen waarom hun aanpak betrouwbaar en valide is. Die onderbouwing is kort, onvolledig en beperkt in kwaliteit (zoals verwacht mag worden van leerlingen van deze leeftijd), maar het inzicht dat die onderbouwing nodig is, is ontwikkeld.

8.1 Samenvatting in het Nederlands

Uit het proefschrift volgt dat leerlingen in klas 4 wel rudimentair kennis hebben van KNO, maar dat die kennis onvoldoende is om in grote mate zelfstandig een degelijk KNO op te zetten en uit te voeren. Ook volgt uit deze onderzoeken dat als we leerlingen willen leren onderzoeken, het laten inzien van het nut van een volledig, juist en onderbouwd antwoord op de onderzoeksvraag een zinvolle strategie is. De ontwikkelde aanpak heeft er voor gezorgd dat er bij leerlingen een cognitieve behoefte is ontstaan om de kennis te ontwikkelen die hen in staat stelt zo'n antwoord te produceren. Daarnaast laat het proefschrift zien dat argumentatie een essentieel onderdeel is van *leren onderzoeken* en dat aandacht voor argumentatie helpt bij het ontwikkelen van onderzoekskennis in leerlingen. De uitkomsten van de verschillende studies dragen bij aan een pedagogisch-didactische theorie voor *leren onderzoeken op basis van argumentatie* waarvoor we in dit proefschrift de basis hebben gelegd. Het algehele promotieonderzoek nodigt uit om verder te onderzoeken hoe het uitgangspunt '*onderzoek doen is het produceren van een wetenschappelijk overtuigende argument voor een claim*' - en daarmee het centraal stellen van argumentatie - gebruikt kan worden om leerlingen te leren onderzoeken.

8.2 Summary in English

Learning to engage in scientific inquiry is an important learning goal of secondary school physics education. Comparable to international curricula, the Dutch physics curriculum states that "the student should be able to carry out an inquiry and draw conclusions using consistent reasoning". However, the literature indicates that this is a seemingly elusive goal. This study therefore aims at determining and understanding how we effectively can develop inquiry knowledge in students in order to enable them to plan, carry out and evaluate a rigorous quantitative physics inquiry (QPI), the type of inquiry often used in education in which the relationship between two quantities is determined. QPI includes many but not all types of physics inquiries. Moreover, being able to engage in scientific inquiry entails more than being able to precisely determine a relationship between two quantities – for instance being able to embed the determined relationship in existing theory. But enabling students to engage in QPI might be a point of departure. If we succeed, we can try to cover other important aspects of inquiry. Important in this is to understand why the chosen approach (opposed to previous attempts) works. If we can explain *why* our approach works, we expand our theoretical knowledge of how to enable students to successfully engage in scientific inquiry.

The first study aims at answering the question what students already know about doing QPI, especially pertaining to analysing empirical data and drawing optimally informative conclusions. On the basis of curriculum documents, we first identify what students entering upper secondary school education ought to know. Using various practicals and related activities, we determine whether the participating students acquired that knowledge. From this study we conclude that the participating 4-HAVO/VWO students (aged ~16) do not attain the expected attainment level. We also find that students do not spontaneously try to produce an informative, correct and substantiated answer to the research question. We argue that students need to be motivated to produce such an answer before they will recognize the problem of their limited attainment level and want to do something about it. Moreover, in teaching scientific inquiry, we argue that not only the four types of knowledge of the *Procedural And Conceptual Knowledge in Science* (PACKS)-model need to be developed in students, it is necessary to integrate argumentation in inquiry. Argumentation plays a crucial role in learning to do science, but remained hitherto underexposed in education.

In the second study we regard scientific inquiry as the construction of a scientifically cogent argument and identify the knowledge a researcher relies on in achieving and assessing that cogency. We regard these insights and views that an experimental researcher relies on in constructing and evaluating scientific evidence, the so-called '*Understandings of Evidence*' (UoE), as important learning goals for engaging students in QPI. As learning goals acquire meaning only if we are able to assess students' attainment level, we specify

conceivable types of actions and decisions expected in inquiry as descriptors for various attainment levels. At the highest level, the students use argumentation to justify their inquiry decisions. The resulting construct consisting of learning goals and descriptors for the various levels, the Assessment Rubric for Physics Inquiry (ARPI), is validated in an augmented Delphi study. The study provides a framework for integrating argumentation and inquiry.

On the basis of this framework and the acquired insights of the first two studies, we build a teaching-learning sequence (TLS) which is described in chapter four and is subject of study in the studies three and four.

To persuade students to invest the effort required to produce a scientifically adequate answer and develop in them the required inquiry knowledge, we must first create in them a need to produce it. In this third study, which focusses on the first activity of the TLS, we try to create this motivation by engaging the students in a contextualized QPI. They evaluate its quality from the consumers' perspective who have to face potentially severe consequences caused by implementing the outcomes of poor research. Considering the risks that are involved within the context, students themselves perceive the quality of their inquiry as insufficient. Students come to understand that in order to produce a meaningful answer, they have to uphold scientific standards. What these standards precisely entail is yet unknown to them but they seem to be willing to develop the associated knowledge.

In the fourth study, this willingness is exploited in three subsequent activities of the TLS where the inquiry knowledge is developed in students. To integrate argumentation in inquiry and foster in students a critical attitude towards their own approach, five design principles were formulated and used in designing these activities. The effectiveness of the design principles is established by verifying whether the expected outcomes are realized. The final activity of the TLS, a contextualized QPI is used to determine what the TLS and a focus on argumentation contributes to student understanding, critical attitude and use of argumentation in doing QPI. By systematically addressing the weaknesses in students' approaches (design principle 4) in a contextualized, guided and basic inquiry (design principle 1-3) and having students express the associated insights and evaluate how these could have been relevant in their first QPI (design principle 5), they acquired selected UoE. As a consequence, in the final QPI students start considering, of their own accord, what scientifically adequate decisions are. The students make a good effort to produce informative conclusions and to substantiate why their approach is reliable and valid. Their substantiation is still brief, incomplete and limited in quality (as can be expected from students of this age), but the insight that such a justification is required has been developed.

From this thesis we learn that students who enter upper secondary education (Grade 10) in the Netherlands have a rudimentary knowledge of doing QPI, but that this knowledge is insufficient to plan and conduct a rigorous QPI independently. The studies also show that if we want to enable students to engage in scientific inquiry, developing in them the insight

that the study should result in a complete, correct and substantiated answer to the research question is a meaningful strategy. Our approach resulted in a cognitive need in students to develop the knowledge that allows them to produce such an answer. In addition, the thesis shows that argumentation is an essential part of scientific inquiry and that explicit attention in teaching to argumentation helps to develop inquiry knowledge in students. The outcomes of the various studies contribute to a pedagogical theory for teaching inquiry through argumentation, for which we have laid the foundation in this thesis. The overall PhD study invites further research into how the idea 'scientific inquiry as the construction of a cogent argument in support of a claim' and thereby putting argumentation at the center can be used to enable students to engage in QPI.

9. Personal reflection

In my study at applied Physics at Delft University of Technology, I felt that I was trained to become an engineer. I never had the feeling that I was trained to become a scientist in the natural sciences. I then became a physics teacher and learned how to engineer education. When I started this research project, I was in no way prepared for the job, I was neither a scientist in the natural sciences or an educational researcher. I did not have the understandings required to independently set up a research project.

At that time I was a part-time researcher and a part-time teacher. As found by others (de Putter-Smits, 2012; van Buuren, 2014), I had to deal with various difficulties stemming from this dual job. First of all, teachers value the practical implications of this thesis. However, scientists especially value the theoretical implications. My focus was, and probably still is, on effective teaching, in helping students understand the world and how it works (although I do not have a definite answer to that question myself). Therefore, what I often wrote at first was relevant for teachers but not publishable in scientific journals.

A second difficulty that I encountered relates to time constraints and deadlines. In teaching the deadline is always tomorrow. You fail when you do not prepare your next lecture properly. In research, the deadline is far, far away. At least in the first few years. Therefore more time was devoted to teaching than probably should have been. The fact that I am not easily satisfied with the quality of my own teaching, further complicated the balance between research and teaching.

As a consequence, cracking this nut (successfully completing this thesis) was a hard and long job. I guess at the end of this research project I really learned to connect the world of the teacher with the world of the researcher. I hope that I came to see and understand how the overarching, in depth studies, provide a theoretical background and insights that can help implement these in science education.

10. Acknowledgement

If a project takes as long as this one, you hardly remember all the people who have provided valuable feedback and support. So, if you are not mentioned here, please be aware that I still will be forever grateful for your support.

First of all, I want to thank my supervisor Peter. I probably was not the easiest student (for you). Stubbornness is one of my virtues, but probably also one of my imperfections (I think you would suggest to discard probably, as almost everything can be probable). In the last two years, I sometimes regretted the topic of argumentation, as many of the comments you provided related to this. In some of the comments you only stated that *an inquiry is the construction of a cogent argument*, implying that an argument was missing or that the text was not convincing (enough). Despite the difficult road we have been walking together, I want to thank you for your stubbornness as it has brought me to where I am now. You have moulded me into the researcher I am today. I sure could not have done it without you.

Second, I want to thank my promotor Marc. Marc, you were there at the moments I needed you the most. You made sure that I continued and regained the motivation required to finish this process.

And now, in random order, I would like to thank various people. The people from ISW Gasthuislaan especially Aad, Freek (and his wife Marieke), Jaap, Monique and Wilfried for their help and support. The people from TUD, especially from SEC and ImPhys with whom I worked together, and Gary for being my mentor in the last part of this process. My friends, Rudi, Forrest, Hendrik, JP, Tom and Patrick for the many discussions we had related to this thesis and your help and willingness to read my work. My family and (other) friends for their (social) support. Especially Rik and Sjoerd, no matter how far or how long we have been apart, meeting you guys always felt convenient.

Laura, I am sorry for the time it took to finish this thesis, for the stress it gave you and me. If obtaining a PhD. is about perseverance, you really deserve one too. I am grateful for your love and support. Lotte and Sofie, you are the joy in my life. I want to thank you for being there and dragging me back to real life. I hope you will hear less frequent ‘not now, dad is busy’. I love the three of you.

11. CV

11.1 CV (NL)

Freek Pols is geboren in Lelystad op 16 april 1986. Na het behalen van zijn VWO-diploma in de Natuur en Techniek profiel in 2004 is hij Technische Natuurkunde gaan studeren aan de Technische Universiteit Delft. Daar behaalde hij zijn master in 2010. Al in 2009, tijdens het afstuderen, begon hij als natuurkunde docent aan het ISW Gasthuislaan in 's-Gravenzande. In 2013 kreeg hij een NWO promotiebeurs voor leraren. Naast zijn docentschap en onderzoek heeft hij veel workshops gegeven, waaronder over *arduino* en *leren onderzoeken*. Ook nam hij actief deel in een professionele leergemeenschap die zich karakteriseert vanuit het idee *van docenten, voor docenten*.

In 2019 werd hij de coördinator van het eerstejaars natuurkunde practicum aan de opleiding Technische Natuurkunde aan de TU Delft. Zijn taak daar is drieledig: het innoveren van het practicumonderwijs naar de 21 -eeuwse onderwijsmaatstaven, de focus leggen op *leren onderzoeken* en de studenten het plezier van het doen van experimentele natuurkunde bijbrengen. Bij deze taak past hij de, in dit onderzoek verworven kennis toe. In november 2022 werd hij een van de TU Delft education fellows: een docent die wordt erkend en gewaardeerd voor zijn/haar inspanningen om onderwijs te innoveren.

11.2 CV (EN)

Freek Pols was born on the 16th of April, 1986 in Lelystad, The Netherlands. He completed the Pre-University Science & Technology track in 2004. He studied applied physics at Delft University of Technology and graduated in 2010. In 2009, when he was still finishing his masters, he started working as a physics teacher at ISW Gasthuislaan, 's-Gravenzande. Besides teaching and research, he gave several workshops, including *Arduino* and *teaching inquiry*. Moreover, he participated in a teacher development community that is characterized by the idea *teachers for teachers.* In 2013 he was granted an NWO scholarship which allowed him to start his PhD study.

In 2019 he started working as the first-year lab course coordinator at the Applied Physics program at the University of Technology, Delft. His task is three fold: innovate the course to the 21st century standards, focus on the development of inquiry knowledge, let students experience the fun of doing experimental physics. He applies the knowledge gained in this study in constantly renewing and innovating the course. In November 2022 he became one of TU Delft's education fellows: a teacher who is recognised and appreciated for his/her efforts for innovating education.

11.3 Peer-reviewed publications

Pols, C.F.J., Diepenbroek, P. (2022). Collaborative Data Collection: Shifting focus on meaning making during practical work. *Physics Education.* In press

Pols, C.F.J. (2022). The Scientific Graphic Organizer for lab work. *The physics teacher.* In press

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2022). Fostering understandings of evidence through development of elements of argumentation. *Under review*

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2022). "Would you dare to jump?" Fostering a scientific approach to secondary physics inquiry. *International Journal of Science Education*, 44(9): 1481-1505

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2021). Defining and Assessing Understandings of Evidence with the Assessment Rubric for Physics Inquiry - Towards Integration of Argumentation and Inquiry. *Physical Review Physics Education Research*, 18(1)

Pols, C.F.J., Duynkerke, L., van Arragon, J., van Prooijen, K., van der Goot, L., & Bera, B. (2021). Students' report on an open inquiry. *Physics Education*, 56(6), 063007.

Pols, C.F.J. (2021). What's inside the pink box? A Nature of Science activity for teachers and students. *Physics Education,* 56(4), 045004

Pols, C.F.J. (2021). The Sound of Music: Determining Young's Modulus using a Guitar String. *Physics Education*, 56 (3), 035027

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2020). What do they know? Investigating students' ability to analyse experimental data in secondary physics education. *International Journal of Science Education*, 43(2): 1-24.

Hut, R.W., Pols, C.F.J., Verschuur, D.J. (2020). Teaching a hands-on course during corona lockdown: from problems to opportunities. *Physics Education* 55 (6), 065022.

Bradbury, F.R., Pols, C.F.J. (2020). A pandemic-resilient open-inquiry physical science lab course which leverages the Maker movement. *The Electronic Journal for Research in Science & Mathematics Education* 24(3).

Pols, C.F.J. (2020). A Physics Lab Course in Times of COVID-19. *The Electronic Journal for Research in Science & Mathematics Education* 24(2): 172-178.

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2019). Introducing argumentation in inquiry—a combination of five exemplary activities. *Physics Education* 54 (5), 055014.

11.4 Conferences

Pols, C.F.J., Dekkers, P.J.J.M. (2022). An introduction to the Assessment Rubric for Physics Inquiry. Paper presented at the GIREP conference, Ljubljana, Slovenia.

Pols, C.F.J., Diepenbroek, P. (2022). The scientific graphic organizer for practical work. Poster presented at the GIREP conference, Ljubljana, Slovenia.

Pols, C.F.J. (2022). Towards inquiry: Redesign of a first year physics lab course. Presentation at the GIREP conference, Ljubljana, Slovenia.

Pols, C.F.J., Lewandowski, H.J., Logman, P.S.W.M., Bradbury, F.R. (2021). Differences and similarities in approaches to physics LAB-courses. Symposium (chair) at the online GIREP World Conference on Physics Education, Hanoi, Vietnam.

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2021). Development of a teaching sequence on physics inquiry. Paper presented at the online GIREP World Conference on Physics Education, Hanoi, Vietnam.

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2021). From producers to consumers: fostering students' scientific approach to practical work. Paper presented at the online ESERA conference, Mino, Braga, Portugal.

Bradbury, F.R., Pols, C.F.J. (2021). Assessing a flipped-lab course consisting of open-inquiry projects using Arduinos. Presentation at AAPT virtual summer meeting.

Bradbury, F.R., Pols, C.F.J. (2021). Using the Assessment Rubric for Physics Inquiry for open inquiries in a multidisciplinary lab course. Poster presentation at Physics Education Research Conference 2021: Making Physics More Inclusive and Eliminating Exclusionary Practices in Physics.

Bradbury, F.R., Pols, C.F.J. (2021). Constraints in Physics Lab Courses: the Good, the Bad, and the Pandemic. Invited speaker at APS March Meeting.

Bradbury, F.R., Pols, C.F.J., Vlaanderen, C.L. (2020). Open-inquiry experiments using sensors controlled by Arduinos in a pandemic-resilient lab course. Poster presentation at Physics Education Research Conference.

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2019). Using cogency to foster the use of concepts of evidence in physics experiments. Paper presented at NARST, Baltimore, USA.

Pols, C.F.J., Dekkers (2018). Data-analysis in practical work, what do students know?. Paper presented at the GIREP conference, San Sebastian, Spain.

Pols, C.F.J. (2017). Enhancing students' data-analysis skills in practical work. Outline study presented at the ESERA summerschool, České Budějovice, Czech Republic.

Pols, C.F.J., Dekkers, P.J.J.M. (2017). *Reality or special effects? Teaching kinematics through debunking film stunts by students.* Workshop given at the ESERA conference, Helsinki, Finland.

Pols, C.F.J. (2015). *Learning kinematics through analysing physics in movies.* Workshop given at the TPI conference, Budapest, Hungary.

11.5 Conference proceedings

Pols, C.F.J., Dekkers, P.J.J.M., de Vries, M.J. (2018). Students' ability to analyse empirical data in practical work. Paper presented at the GIREP conference, San Sebastian, Spain. Journal of Physics: Conference Series 1287 (1), 012001. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1287/1/012001/pdf

Pols, C.F.J. (2015). Real or fake? What can students learn from debunking Hollywood physics? Workshop presented at Teaching Physics Innovatively, Budapest, Hungary. Available at: http://real.mtak.hu/73483/2/tpi_bel_kuldesre1.pdf

11.6 Related publications

Pols, C.F.J. (2022). A taste of history (original title: Proeven (v/a)an vroeger), *NVOX* 47 (10), 286-286.

Pols, C.F.J., Hut, R. W. Maker Education in the Applied Physics Bachelor Programme at Delft University of Technology. Book. M.J. de Vries (ed). Under review.

Pols, C.F.J., (2021). Practicals on the topic gravitational acceleration (original title: Practica bij de valbeweging). *NVOX* 46 (10), 34-35.

Pols, C.F.J., (2021). Critical? Let them experience... Development of a critical attitude in a teaching sequence on scientific inquiry (original title: Kritisch? Laat het ze maar ervaren Ontwikkeling van een kritische houding in een leerlijn onderzoeken.) *NTvN* 87 (8), 10-13.

Pols, C.F.J., Hut, R.W., Oosterlaan, L., van Braak, M., Collenteur, F., (2020). Build a demonstrationexperiment (original title: Bouw een demoproef). *NVOX* 45 (10), 58-59.

Pols, C.F.J. (2020). Practicals, just a little different (original title: Practicum, net even anders). *NVOX* 45 (8), 438-439.

Pols, C.F.J. (2019). The Scientific Graphic Organizer (original title: De Scientific Graphic Organizer). *NVOX* 44 (8), 410-411.

Pols, C.F.J. (2019). The Vitruvian Man (original title: De mens van Vetruvius), *NVOX* 44 (6), 286-286.

Pols, C.F.J. (2019). Teaching inquiry: a practical approach in grade 10 (original title: Leren onderzoeken: een praktische aanpak in klas 4). *NVOX* 44 (2), 98-99.

Pols, C.F.J. (2018). Teaching inquiry to secondary school students (original title: Leerlingen leren onderzoeken). *Nederlands Tijdschrift voor Natuurkunde,* 84 (11), 45-47.

Pols, C.F.J. (2018). The slope test: A practical on resolving forces in components (original title: De hellingproef. Een practicum over het ontbinden van krachten in componenten). *NVOX,* 47 (8), 442-443.

Pols, C.F.J. (2018). Motion in stop-motion (original title: Beweging in stop-motion). *NVOX,* 43 (5), 244-245.

Pols, C.F.J. (2017). String theory in practice (original title: Snaartheorie in de praktijk). *NVOX,* 42 (2), 72-73.

Pols, C.F.J. (2017). Engineering design in grade 10 (original title: Technisch ontwerpen in 4V). *NVOX,* 42 (1), 40-41.

Frederik, I., van den Berg, E., te Brinke, L., Dekkers, P., Pols, F., Sonneveld, W., Spaan, W., van Veen, N., van Woerkom, M., (2017). Show de Fysica 2: Natuurkunde laat je zien. NVON: Nederlandse Vereniging voor Onderwijs in de Natuurwetenschappen

Mooldijk, A. Pols, C.F.J. (2017). Leerstofdomeinen: Technische automatisering in Kortland, K., Mooldijk, A. Poorthuis, H., Handboek natuurkundedidactiek, 206-211.

Pols, C.F.J. (2017). Accelerating down a slope (original title: Versnelling langs een helling). *NVOX,* 40 (4), 178-179.

Pols, C.F.J. (2016). A picture of graph says a 1000 words (original title: Een foto of grafiek zegt meer dan 1000 woorden). *NVOX,* 41 (2), 68-69.

11.7 Relevant workshops

Pols, C.F.J. (2020). *Didactics of practical work.* Teacher-trainer training given for the Hogeschool Utrecht, Utrecht, The Netherlands.

Pols, C.F.J. (2019). *Scientific Inquiry in STEM Education.* Teacher training given for the Croatian science teacher network, Split, Croatia.

Pols, C.F.J. (2019). *A teaching sequence on scientific inquiry*. Teacher training given at the WND-conference, Noordwijkerhout, The Netherlands.

Pols, C.F.J. (2018). *A teaching sequence on scientific inquiry*. Teacher training given at the WND-conference, Noordwijkerhout, The Netherlands.

Pols, C.F.J. (2018). *Didactics of practical work*. Teacher training given at the Jong meeting, Delft, The Netherlands.

Pols, C.F.J. (2017). *The development of inquiry skills*. Teacher training given at the WND-conference, Noordwijkerhout, The Netherlands.

12.Appendix

12.1 List of abbreviations

Abbr.	abbreviation	explanation
AAPT	American Association of Physics Teachers	A professional membership association of scientists dedicated to enhancing the understanding and appreciation of physics through teaching.
ARPI	Assessment Rubric for Physics Inquiry	An rubric for assessing students' attainment level of UoE in QPI developed by Pols et al. (2022a).
CoE	Concepts of Evidence	A tentative list of ~100 concepts that underpin the concepts reliability and validity developed by (Gott & Duggan, 1996)
ECLASS	Colorado Learning Attitudes about Science Survey for Experimental Physics	An assessment tool (survey) for undergraduate physics lab course developed by Zwickl, Finkelstein, and Lewandowski (2013).
EDR	Educational Design Research	A research approach that aims at developing theoretical insights and practical solutions through a combined study of both the process of learning and the means that support that process (Van den Akker et al., 2006).
ISL	International Swimming League	A fictitious organization assessing the fairness of swimming competition.
NGSS	Next Generation Science Standards	A framework for K–12 science education from which curricula can be developed.
NRC	National Research Council	The operating arm of the United States National Academies of Sciences, Engineering, and Medicine. This organization developed the framework on which the NGSS are based.
OECD	Organisation for Economic Co-operation and Development	A global policy forum of countries working together to improve the economic and social well-being of people. This organisation developed the PISA
PACKS	Procedural and Conceptual Knowledge in Science	A model developed by Millar et al. (1994) to link various types of knowledge to decisions made in various stages of an inquiry.
PISA	Programme for International Student Assessment	The PISA is geared towards assessing scientific literacy internationally. PISA measures 15-year-olds' ability to use their reading, mathematics and science knowledge and skills to meet real-life challenges.
PLIC	Physics Lab Inventory of Critical thinking	A closed response survey designed to assess how students critically evaluate experimental methods, data, and models (Walsh et al., 2019).
QPI	Quantitative Physics Inquiry	The type of physics inquiry in which a mathematical relationship between two quantities is to be determined.

SGO	Scientific Graphic Organiser	A written pre-structured lab journal providing a schematic for reporting the essentials of an inquiry (Pols, 2019).
TAs	Teaching assistants	Experience students supporting the teacher in running a lab course.
TLS	Teaching-learning sequence	A series of educational activities aimed at
UoE	Understandings of Evidence	UoE express insights, principles and procedures an experimental researcher relies on in constructing, presenting and evaluating scientific evidence for QPI.

12.2 Appendix related to chapter 2

Probe I1:

Q1a: How would you describe the graph to someone who does not see the graph?
Q1b: Have you described all essential features, or would you like to add something?

Probe I2:

Q1: What kind of event can this be a graph of?
Q2a: How would you describe the graph to someone who does not see the graph?
Q2b: Are there any special data points?
Q3a: What would be the line that best fits these measurements? Discuss and subsequently draw that line.
Q3b: Are there other lines that could fit as well?
Q4: Would a line going through the origin fit well with these measurements?
Q5: What conclusion can you draw from this graph about the motion described with this graph?
Q6: After how much time has the ball travelled 4 meters?
Q7: After how much time has the ball travelled 9 meters?

Probe I3:

Q1: What could these measurements be about?
Q2: What would be the line that best fits these measurements? Discuss and subsequently draw that line.
Q3: How would you describe the graph to someone who does not see the graph?
Q4a: Is the description you just gave the same as you would write in a report?
Q4b: Why (not)? What else would you include? Why?

Q5: Using these measurements, what could a conclusion be about?

Q6: What additional information would you like to receive in order to draw a more reliable conclusion?

Q7: What was the most difficult part of the interview?

13.References

Abrahams, I. (2005). *Between rhetoric and reality: The Use and effectiveness of practical work in secondary school science.* (PhD). University of York, UK,

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

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

Abrahams, I., & Reiss, M. J. (2012). Practical work: Its effectiveness in primary and secondary schools in England. *Journal of Research in science teaching, 49*(8), 1035-1055. doi:10.1002/tea.21036

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

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

Aikenhead, G. S. (2005). Science-based occupations and the science curriculum: Concepts of evidence. *Science Education, 89*(2), 242-275.

Allen, S., & Knight, J. (2009). A Method for Collaboratively Developing and Validating a Rubric. *International Journal for the Scholarship of Teaching and Learning, 3*(2), n2.

Allie, S., Buffler, A., Campbell, B., & Lubben, F. (1998). First-year physics students' perceptions of the quality of experimental measurements. *International Journal of Science Education, 20*(4), 447-459.

Altricher, H., Feldman, A., Posch, P., & Somekh, B. (2005). *Teachers investigate their work: An introduction to action research across the professions*. Abingdon, Oxo: Routledge.

American Psychological Association. (1983). *Publication manual*: American Psychological Association Washington, DC.

Andersson, B., & Bach, F. (2005). Developing new teaching sequences in science: the example of 'Gases and their properties'. In *Research in science education in europe* (pp. 13-25): Routledge.

Ansell, K., & Selen, M. (2016). *Student attitudes in a new hybrid design-based introductory physics laboratory*. Paper presented at the Physics Education Research Conference, Sacramento, CA.

Bailey, S., & Millar, R. (1996). From logical reasoning to scientific reasoning: students' interpretation of data from science investigations. *Science Education Research Paper, 96*(01).

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

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

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The journal of the learning sciences, 13*(1), 1-14.

Barron, B. J., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A., Zech, L., & Bransford, J. D. (1998). Doing with understanding: Lessons from research on problem-and project-based learning. *Journal of the learning sciences, 7*(3-4), 271-311.

Bell, P., Hoadley, C. M., & Linn, M. C. (2004). Design-based research in education. *Internet environments for science education, 2004*, 73-85.
Bell, R. L., Smetana, L., & Binns, I. (2005). Simplifying inquiry instruction. *The Science Teacher, 72*(7), 30-33.
Bennett, J., Lubben, F., & Hogarth, S. (2007). Bringing science to life: A synthesis of the research evidence on the effects of context-based and STS approaches to science teaching. *Science Education, 91*(3), 347-370. doi:10.1002/sce.20186
Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347-364.
Black, P., & Wiliam, D. (2005). *Inside the black box: Raising standards through classroom assessment*: Granada Learning.
Boeker, E., & Van Grondelle, R. (2011). *Environmental physics: sustainable energy and climate change*: John Wiley & Sons.
Boohan, R. (2016a). The language of mathematics in science. *School science review, 97*, 15-20.
Boohan, R. (2016b). The Language of Mathematics in Science: A Guide for Teachers of 11–16 Science. In: Hatfield: Association for Science Education. Available at: www. ase. org. uk/resources/maths-in-science.
Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance.
Brewer, M. B. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3-16). Cambridge: Cambridge University Press.
Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment. The Missing Part of Pedagogy. ASHE-ERIC Higher Education Report* (Vol. 27). Washington DC: ERIC.
Bruckheimer, J. (Producer). (2007). Pirates of the Caribbean: at world's end. *Pirates of the Caribbean*. Retrieved from https://youtu.be/SfyePrFKvVA
Bryman, A. (2015). *Social research methods*. Oxford: Oxford university press.
Buffler, A., Allie, S., & Lubben, F. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education, 23*(11), 1137-1156.
Burdett, N., & Sturman, L. (2013). *A Comparison of PISA and TIMSS against England's National Curriculum.* Paper presented at the 5th IEA International Research Conference–26-28 June.
Carr, W., & Kemmis, S. (2003). *Becoming critical: education knowledge and action research*. Abingdon, Oxo: Routledge.
Cavagnetto, A. R. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts. *Review of educational research, 80*(3), 336-371.
Chalmers, A. F. (2013). *What is this thing called science?* : Hackett Publishing.
Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education*: routledge.
Crawford, B. A. (2014). From Inquiry to Scientific Practices in the Science Classroom. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. 2, pp. 515-541): Routledge.
Custer, R. L., Scarcella, J. A., & Stewart, B. R. (1999). The modified Delphi technique-A rotational modification. *Journal of Vocational and Technical Education, 15*(2).
Dancy, M., Lau, A. C., Rundquist, A., & Henderson, C. (2019). Faculty online learning communities: A model for sustained teaching transformation. *Physical Review Physics Education Research, 15*(2), 020147.
de Putter-Smits, L. G. A. (2012). *Science teachers designing context-based curriculum materials: developing context-based teaching competence.* PhD thesis). doi: 10.6100/IR724553,
Dehn, M. J. (2011). *Working memory and academic learning: Assessment and intervention*: John Wiley & Sons.

Dekkers, P. J. J. M. (1997). *Making productive use of Students Conceptions in Physics Education: Developing the Concept of Force through Practical Work.* (phd). Vrije Universiteit Amsterdam, Amsterdam.
Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning: A guide to nominal group and Delphi processes*: Scott, Foresman.
Department for Education England. (2013). Science programmes of study: key stage 3. National curriculum in England. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/335174/SECONDARY_national_curriculum_-_Science_220714.pdf
Dillon, J. (2008). A Review of the Research on Practical Work in School Science. *King's College, London*, 1-9.
DiSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *The journal of the learning sciences, 13*(1), 77-103.
Dounas-Frazer, D. R., & Lewandowski, H. (2018). The modelling framework for experimental physics: Description, development, and applications. *EUROPEAN JOURNAL OF PHYSICS, 39*(6), 064005.
Driver, R. (1995). Constructivist approaches to science teaching. *Constructivism in education*, 385-400.
Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287-312. doi:10.1002/(SICI)1098-237X(200005)84
DUO. (2017). Programs in secondary education.
Durant, J. (1994). What is scientific literacy? *European Review, 2*(1), 83-89.
Duschl, R. (2000). Making the nature of science explicit. In R. Millar, J. Leach, & J. Osborne (Eds.), *Improving science education: The contribution of research* (pp. 187-206). London: Buckingham: Open University Press.
Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of research in education, 32*(1), 268-291.
Duschl, R., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education, 38*(1), 39-72. doi:10.1080/03057260208560187
Eijkelhof, H. (2014). *Curriculum policy implications of the pisa scientific literacy framework.* Paper presented at the Electronic Proceedings of the ESERA 2013 Conference, Strand 10, Science Curriculum and Educational Policy, Nicosia, Cyprus.
Eijkelhof, H., & Kortland, J. (2001). Bouwen aan een nieuw natuurkundecurriculum havo/vwo. *Tijdschrift voor Didactiek der Betawetenschappen, 18*(1), 2-18.
Erduran, S. (2018). Toulmin's argument pattern as a "horizon of possibilities" in the study of argumentation in science education. *Cultural Studies of Science Education*, 1-9.
Erduran, S., & Jiménez-Aleixandre, M. P. (2008). Argumentation in science education. *Perspectives from classroom-Based Research. Dordrecht: Springer*.
Erduran, S., Osborne, J., & Simon, S. (2005). The role of argumentation in developing scientific literacy. In *Research and the quality of science education* (pp. 381-394): Springer.
Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915-933.
Etkina, E. (2015). Millikan award lecture: Students of physics—Listeners, observers, or collaborative participants in physics scientific practices? In: AAPT.
Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D. T., Gentile, M., Murthy, S., . . . Warren, A. (2006). Scientific abilities and their assessment. *Physical Review Special Topics-Physics Education Research, 2*(2), 020103.
European Commission. (1995). *White paper on education and training. Teaching and Learning: Towards the learning society*. Brussels Retrieved from

https://op.europa.eu/nl/publication-detail/-/publication/d0a8aa7a-5311-4eee-904c-98fa541108d8/language-en
Farmer, S. (2012). Real Graphs from Real Data: Experiencing the Concepts of Measurement and Uncertainty. *School science review, 346*, 81-84.
Feynman, R. P., Leighton, R. B., & Sands, M. (2011). *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat* (Vol. 1): Basic books.
Fitch, J. C. (1971). Energy absorbing deceleration barriers. In: Google Patents.
Gast, D. L. (2014). General factors in measurement and evaluation. In *Single case research methodology* (pp. 85-104): Routledge.
Giddings, G. J., Hofstein, A., & Lunetta, V. N. (1991). Assessment and evaluation in the science laboratory. In B. E. Woolnough (Ed.), *Practical science* (pp. 167-178). Milton Keynes - Philadelphia: Open University Press.
Gilbert, J. K. (2006). On the nature of "context" in chemical education. *International Journal of Science Education, 28*(9), 957-976.
Glaesser, J., Gott, R., Roberts, R., & Cooper, B. (2009). Underlying success in open-ended investigations in science: using qualitative comparative analysis to identify necessary and sufficient conditions. *Research in science & technological education, 27*(1), 5-30. doi:10.1080/02635140802658784
Gott, R., & Duggan, S. (1995). *Investigative Work in the Science Curriculum. Developing Science and Technology Education*. Buckingham, England: Open University Press.
Gott, R., & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education, 18*(7), 791-806. doi:10.1080/0950069960180705
Gott, R., & Duggan, S. (2003). *Understanding and using scientific evidence: How to critically evaluate data*. Buckingham: Sage Publications Ltd.
Gott, R., & Duggan, S. (2007). A framework for practical work in science and scientific literacy through argumentation. *Research in science & technological education, 25*(3), 271-291. doi:10.1080/02635140701535000
Gott, R., Duggan, S., Roberts, R., & Hussain, A. (2003, 2018). Research into understanding scientific evidence. Retrieved from http://www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm
Gott, R., & Roberts, R. (2008). 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*.
Guetterman, T. C. (2015). *Descriptions of sampling practices within five approaches to qualitative research in education and the health sciences.* Paper presented at the Forum Qualitative Sozialforschung/Forum: Qualitative Social Research.
Gunstone, R. F., & Champagne, A. B. (1990). Promoting conceptual change in the laboratory. In E. H. Hazel (Ed.), *The student laboratory and the science curriculum* (pp. 159-182): Routledge.
Gurria, A. (2016). PISA 2015 results in focus. *PISA in Focus*(67), 1.
Hart, C., Mulhall, P., Berry, A., Loughran, J., & Gunstone, R. (2000). What is the purpose of this experiment? Or can students learn something from doing experiments? *Journal of Research in science teaching, 37*(7), 655-675.
Henderson, T. (1996). https://www.physicsclassroom.com/The-Laboratory.
Hihi, J. D. (2020). SPOON: Reinvest shared activities during meetings. Retrieved from https://spoonobook.hypotheses.org/files/2019/12/SPOON-activity-reinvest-swedish-road-service-.docx
Hodson, D. (1988). Experiments in science and science teaching. *Educational philosophy and theory, 20*(2), 53-66.
Hodson, D. (1990). A critical look at practical work in school science. *School science review, 70*(256), 33-40. Retrieved from https://eric.ed.gov/?id=EJ413966

Hodson, D. (1991). Practical work in science: Time for a reappraisal. In E. Hegarty-Hazel (Ed.), *Studies in Science Education*. London: Routledge.
Hodson, D. (1992). Assessment of practical work. *Science & Education, 1*(2), 115-144. doi:10.1007/BF00572835
Hodson, D. (1993). Re-thinking old ways: Towards a more critical approach to practical work in school science. *Studies in Science Education, 22*, 85– 142.
Hodson, D. (1994). Redefining and reorienting practical work in school science. *Teaching science*, 159-163.
Hodson, D. (2001). *Research on practical work in school and universities: In pursuit of better questions and better methods.* Paper presented at the Proceedings of the 6th European Conference on Research in Chemical Education, University of Aveiro, Aviero, Portugal.
Hodson, D. (2014). Learning science, learning about science, doing science: Different goals demand different learning methods. *International Journal of Science Education, 36*(15), 2534-2553. doi:10.1080/09500693.2014.899722
Hofstein, A. (2017). The role of laboratory in science teaching and learning. In K. S. Taber & B. Akpan (Eds.), *Science Education* (pp. 357-368). Dordrecht: Springer.
Hofstein, A., & Kind, P. M. (2012). Learning in and from science laboratories. In B. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 189-207). Dordrecht, The Netherlands: Springer.
Hofstein, A., & Lunetta, V. N. (1982). The role of the laboratory in science teaching: Neglected aspects of research. *Review of educational research, 52*(2), 201-217.
Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education, 88*(1), 28-54. doi:10.1002/sce.10106
Holmes, N. G., Olsen, J., Thomas, J. L., & Wieman, C. (2017). Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content. *Physical Review Physics Education Research, 13*(1), 010129.
Holmes, N. G., & Wieman, C. (2016). Examining and contrasting the cognitive activities engaged in undergraduate research experiences and lab courses. *Physical Review Physics Education Research, 12*(2), 020103. doi:10.1103/PhysRevPhysEducRes.12.020103
Holmes, N. G., & Wieman, C. (2018). Introductory physics labs: We can do better. *Physics Today, 71*, 1-38. doi:10.1063/PT.3.3816
Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical Assessment, Research, and Evaluation, 12*(1), 10. doi:https://doi.org/10.7275/pdz9-th90
Hurd, P. D. (1998). Scientific literacy: New minds for a changing world. *Science Education, 82*(3), 407-416.
Jenkins, E. (1998). The schooling of laboratory science. *Practical work in school science: Which way now*, 35-51.
Johnson, S., Britain, G., & Unit, S. A. o. P. (1989). *National Assessment: the APU science approach*. Lodon: HM Stationery Office.
Johnstone, A. H., & Wham, A. (1982). The demands of practical work. *Education in chemistry, 19*(3), 71-73.
Jones, L. R., Wheeler, G., & Centurino, V. A. (2015). TIMSS 2015 science framework. In *TIMMS* (pp. 29-58).
Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in science teaching, 41*(7), 748-769. doi:10.1002/tea.20020
Kapur, M. (2008). Productive failure. *Cognition and instruction, 26*(3), 379-424.
Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561-579.

Kelly, G. J. (2014). Discourse practices in science learning and teaching. In N. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. 2, pp. 321-336).
Kempa, R. (1986). *Assessment in science*: Cambridge University Press.
Keogh, B., & Naylor, S. J. I. J. o. S. E. (1999). Concept cartoons, teaching and learning in science: an evaluation. *21*(4), 431-446.
Kim, H., & Song, J. (2006). The features of peer argumentation in middle school students' scientific inquiry. *Research in Science Education, 36*(3), 211-233.
Kind, P. (1999). TIMSS Performance Assessment-a cross national comparison of practical work. *Practical Work in Science Education–Recent Research Studies*, 75-95.
Kok, K., Priemer, B., Musold, W., & Masnick, A. (2019). Students' conclusions from measurement data: The more decimal places, the better? *Phys. Rev. Phys. Educ. Res., 15*(1).
Kortland, J. (2007). *Context-based science curricula: Exploring the didactical friction between context and science content.* Paper presented at the ESERA 2007 Conference, Malmö, Sweden. To be retrieved from the author's website: www. phys. uu. nl/~ kortland> English> Publications.
Kozminski, J., Lewandowski, H., Beverly, N., Lindaas, S., Deardorff, D., Reagan, A., . . . Hobbs, R. (2014). *AAPT recommendations for the undergraduate physics laboratory curriculum*. Retrieved from
Kratochwill, T. R. (2013). *Single subject research: Strategies for evaluating change*: Academic Press.
Lachmayer, S., Nerdel, C., & Prechtl, H. (2007). Modelling of cognitive abilities regarding the handling of graphs in science education. *Zeitschrift für Didaktik der Naturwissenschaften, 13*, 161-180.
Larkin, J. H., & Reif, F. (1979). Understanding and teaching problem-solving in physics. *European journal of science education, 1*(2), 191-203.
Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science Education, 84*(1), 71-94.
Lavoie, J.-M., & Montpetit, R. R. (1986). Applied physiology of swimming. *Sports medicine, 3*(3), 165-189.
Lederman, J. S., Lederman, N. G., Bartos, S. A., Bartels, S. L., Meyer, A. A., & Schwartz, R. S. J. J. o. r. i. s. t. (2014). Meaningful assessment of learners' understandings about scientific inquiry—The views about scientific inquiry (VASI) questionnaire. *Journal of Research in science teaching, 51*(1), 65-83.
Lederman, N. G., & Abd-El-Khalick, F. (1998). Avoiding de-natured science: Activities that promote understandings of the nature of science. In *The nature of science in science education* (pp. 83-126): Springer.
Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of educational research, 60*(1), 1-64.
Lewandowski, H. (2014). Colorado Learning about Science Survey for Experimental Physics (E-CLASS). *APS, 2014*, S38. 003.
Lijnse, P. (2014). *Omzien in verwarring*. Utrecht, The Netherlands: Fisme.
Lijnse, P. L. (1995). "Developmental research" as a way to an empirically based "didactical structure" of science. *Science Education, 79*(2), 189-199.
Lipton, P. (2003). *Inference to the best explanation*. Oxfordshire, England,: Routledge.
Livingston, J. A. (2003). Metacognition: An Overview.
Loren, A. A. (1992). *ASSURED CREW RETURN VEHICLE.* Paper presented at the UNIVERSITY ADVANCED DESIGN PROGRAM, Washington D.C.
Lubben, F., Campbell, B., Buffler, A., & Allie, S. (2001). Point and set reasoning in practical science measurement by entering university freshmen. *Science Education, 85*(4), 311-327.
Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education, 18*(8), 955-968.

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

Mason, J. (2002). *Researching your own practice: The discipline of noticing*: Routledge.

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

McDermott, L. C., & Redish, E. F. (1999). Resource letter: PER-1: Physics education research. *American Journal of Physics, 67*(9), 755-767.

McKenney, S., & Reeves, T. C. (2013). *Conducting educational design research*. Abingdon, Oxo: Routledge.

McNamara, T. F., & Macnamara, T. J. (1996). *Measuring second language performance*. London: Longman Publishing Group.

Méheut, M., & Psillos, D. (2004). Teaching–learning sequences: aims and tools for science education research. *International Journal of Science Education, 26*(5), 515-535. doi:10.1080/09500690310001614762

Millar, R. (1991). A means to an end: The role of processes in science education. *Practical science*, 43-52.

Millar, R. (1997). Student's understanding of the procedures of scientific enquiry. In A. Tiberghien, E. L. Jossem, & J. Barojas (Eds.), *Connecting Research in Physics Education with Teacher Education* (pp. 65-70): International Commission on Physics Education.

Millar, R. (1998). Rhetoric and reality: What practical work in science education is really for. *Practical work in school science: Which way now*, 16-31.

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

Millar, R. (2008). *Taking scientific literacy seriously as a curriculum aim.* Paper presented at the Asia-Pacific Forum on Science Learning and Teaching, Hong Kong.

Millar, R. (2009). Analysing practical activities to assess and improve effectiveness: The Practical Activity Analysis Inventory (PAAI). *York: Centre for Innovation and Research in Science Education, University of York*.

Millar, R. (2010). Practical work. In J. Osborne & J. Dillon (Eds.), *Good Practice In Science Teaching: What Research Has To Say: What research has to say* (2nd ed., pp. 108): Open University Press.

Millar, R. (2015). Experiments. *Encyclopaedia of science education*, 418-419.

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

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

Millar, R., & Osborne, J. (1998). *Beyond 2000: Science education for the future*. Retrieved from London: https://www.nuffieldfoundation.org/wp-content/uploads/2015/11/Beyond-2000.pdf

Miller, L. (2006). *Determining what could/should be: The Delphi technique and its application.* Paper presented at the meeting of the 2006 annual meeting of the Mid-Western Educational Research Association, , Columbus, Ohio.

Ministry of Education Singapore. (2013). *Science Syllabus Lower and Upper Secondary*. Retrieved from https://www.moe.gov.sg/docs/default-

source/document/education/syllabuses/sciences/files/science-lower-upper-secondary-2014.pdf
Molyneux-Hodgson, S., Sutherland, R., & Butterfield, A. (1999). Is 'Authentic'Appropriate? The Use of Work Contexts in Science Practical Activity. In J. Leach & A. Paulsen (Eds.), *Practical Work in Science Education: Recent Research Studies* (pp. 160-174). Alphen aan den Rijn, Netherlands: Kluwer
Mooldijk, A., & Savelsbergh, E. (2000). An example of the integration of modeling into the curriculum: a falling cone. *Proceedings of the GIREP: Physics Teacher Education Beyond*, 625-628.
Mooldijk, A., & Sonneveld, W. (2010). *Coherent education in mathematics and physics: the theme of proportionality in mathematics and physics*. Paper presented at the Trend in Science and Mathematics Education (TiSME).
Mooldijk, A., van der Valk, T., & Wooning, J. (2006). Top Angle and the Maximum Speed of Falling Cones. *Science Education International, 17*(3), 161-169.
Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research, and Evaluation, 7*(1), 3.
Murry Jr, J. W., & Hammons, J. O. (1995). Delphi: A versatile methodology for conducting qualitative research. *The review of higher education, 18*(4), 423-436.
Najami, N., Hugerat, M., Kabya, F., & Hofstein, A. (2020). The Laboratory as a Vehicle for Enhancing Argumentation Among Pre-Service Science Teachers. *Science & Education*, 1-17. doi:10.1007/s11191-020-00107-9
National Research Council. (2000). *Inquiry and the National Science Education Standards: A Guide for Teaching and Learning*. Washington, DC: The National Academies Press.
Netherlands Institute for Curriculum Development. (2016). *Retrieved from http://international.slo.nl*.
Newton, P., Driver, R., & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education, 21*(5), 553-576. doi:10.1080/095006999290570
Next Generation Science Standards. (2013). Next generation science standards: For states, by states. *Appendix D: All standards, all students: Making the Next Generation Science Standards accessible to all students*.
NRC. (2013). Next generation science standards: For states, by states.
Ntombela, G. (1999). A marriage of inconvenience? School science practical work and the nature of science. In J. Leach & A. C. Paulsen (Eds.), *Practical Work in Science Education: Recent Research Studies* (pp. 118-133).
OECD. (2013). *PISA 2015: DRAFT SCIENCE FRAMEWORK*. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm
Oreskes, N. (2018). The scientific consensus on climate change: How do we know we're not wrong? In *Climate Modelling* (pp. 31-64): Springer.
Oreskes, N. (2019). *Why Trust Science?* (Vol. 1): Princeton University Press.
Osborne, J. (2000). Science for citizenship. *Good practice in science teaching: What research has to say*, 225-240.
Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265-279. doi:10.1016/j.tsc.2013.07.006
Osborne, J. (2014a). Scientific practices and inquiry in the science classroom. In N. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. 2, pp. 579 - 599).
Osborne, J. (2014b). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*(2), 177-196.
Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections* (Vol. 13): London: The Nuffield Foundation.

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

Partnership, G. S. The glossary of educational reform. Retrieved from https://www.edglossary.org/portfolio/

Pfeffer, J., & Sutton, R. I. (1999). *The knowing-doing gap: How smart companies turn knowledge into action*. Boston, Massachusetts: Harvard business press.

Phys.org. (2011). Siberian region 'confirms Yeti exists'. Retrieved from https://phys.org/news/2011-10-siberian-region-yeti.html

Pols, C. F. J. (2016). Een foto of grafiek zegt meer dan 1000 woorden (A graph explains more than a 1000 words). *NVOX*(2), 2.

Pols, C. F. J. (2017). Snaartheorie in de praktijk (String theory in practice). *NVOX, 42*(2), 2.

Pols, C. F. J. (2019). De Scientific Graphic Organizer. *NVOX, 44*(8), 410-411.

Pols, C. F. J. (2020a). A Physics Lab Course in Times of COVID-19. *The Electronic Journal for Research in Science & Mathematics Education, 24*(2), 172-178.

Pols, C. F. J. (2020b). Practicum, net even anders (Practicals, just a little different). *NVOX, 48*(8), 438-439.

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

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

Pols, C. F. J., Dekkers, P. J. J. M., & de Vries, M. J. (2019). Introducing argumentation in inquiry—a combination of five exemplary activities. *Physics education, 54*(5), 055014. doi:10.1088/1361-6552/ab2ae5

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

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

Pols, C. F. J., Dekkers, P. J. J. M., & de Vries, M. J. (2022b). "Would you dare to jump?" Fostering a scientific approach to research in secondary physics education. *International Journal of Science Education, 44*(7), 1481-1505 doi:https://doi.org/10.1080/09500693.2022.2083251

Pospiech, G., Geyer, M., Ceuppens, S., De Cock, M., Deprez, J., Dehaene, W., . . . Stefanel, A. (2019). *Role of graphs in the mathematization process in physics education.* Paper presented at the Journal of Physics: Conference Series, San-Sebastian.

Redish, E. F., & Rigden, J. S. (1998). *The Changing Role of Physics Departments in Modern Universities: Proceedings of ICUPE*: AIP Press [Imprint].

Roberts, D. A., & Bybee, R. W. (2014). Scientific literacy, science literacy, and science education. In N. Lederman & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. 2). New York: Routledge.

Roberts, R., Gott, R., & Glaesser, J. (2010). Students' approaches to open-ended science investigation: the importance of substantive and procedural understanding. *Research Papers in Education, 25*(4), 377-407.

Roberts, R., & Johnson, P. (2015). Understanding the quality of data: a concept map for 'the thinking behind the doing'in scientific practice. *The Curriculum Journal, 26*(3), 345-369.

Roberts, R., & Reading, C. (2015). The practical work challenge: incorporating the explicit teaching of evidence in subject content. *School science review.*(357), 31-39.

Roll, I., Holmes, N. G., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. *Instructional Science, 40*(4), 691-710.

Roorda, G., Vos, P., & Goedhart, M. J. (2015). An actor-oriented transfer perspective on high school students'development of the use of procedures to solve problems on rate of change. *International Journal of Science and Mathematics Education, 13*(4), 863-889. doi:10.1007/s10763-013-9501-1

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social work research, 27*(2), 94-104.

Rusman, E., & Dirkx, K. (2017). Developing Rubrics to Assess Complex (Generic) Skills in the Classroom: How to Distinguish Skills' Mastery Levels? *Practical Assessment, Research, and Evaluation, 22*(1), 12. doi:https://doi.org/10.7275/0eat-hb38

Sampson, V., Grooms, J., & Walker, J. P. (2009). Argument-driven inquiry. *The Science Teacher, 76*(8), 42.

Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments: An exploratory study. *Science Education, 95*(2), 217-257.

Schalk, H. H., Van der Schee, J. A., & Boersma, K. T. (2008, September). *The use of concepts of evidence by students in biology investigations: Development research in pre-university education.* Paper presented at the 7th ERIDOB Conference (págs. 1-12). Netherlands: Utrecht University.

Schwartz, R. S., Lederman, N. G., & Crawford, B. A. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education, 88*(4), 610-645.

Séré, M. G., Journeaux, R., & Larcher, C. (1993). Learning the statistical analysis of measurement errors. *International Journal of Science Education, 15*(4), 427-438.

Singapore, M. o. E. (2019). *Physics Syllabus Pre-University*. Retrieved from https://www.moe.gov.sg/docs/default-source/document/education/syllabuses/sciences/files/preuniversity_h2_physics_syllabus.pdf

Smits, T., Lijnse, P., & Bergen, T. (2000). Leerlingonderzoek met kwaliteit. *Tijdschrift voor Didactiek der B-wetenschappen, 17*(1).

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

Spaan, W., Oostdam, R., Schuitema, J., & Pijls, M. (2022). Analysing teacher behaviour in synthesizing hands-on and minds-on during practical work. *Research in Science and Technological Education, 40*, 19. doi:10.1080/02635143.2022.2098265

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

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

Stump, E. M., White, C. L., Passante, G., & Holmes, N. (2020). Student reasoning about sources of experimental measurement uncertainty in quantum versus classical mechanics. *arXiv preprint arXiv:2007.06675*.

Sunder, S. G. (2016). *Connecting IB to the NGSS: The Dual Implementation of International Baccalaureate and the Next Generation Science Standards:Challenges and Opportunities*. Retrieved from Geneva:

https://static1.squarespace.com/static/59c3bad759cc68f757a465a3/t/5aad1b0c352f533ca53a77d1/1521294147614/IB+and+NGSS.pdf
Tamir, P. (1991). Practical work in school science: an analysis of current practice. In B. E. Woolnough (Ed.), *Practical science* (pp. 13-20). Milton Keynes - Philadelphia: Open University press.
Tasker, R., & Freyberg, P. (1985). Facing the mismatches in the classroom. *Learning in science: The implications of children's science*, 66-80.
Toulmin, S. E. (2003). *The uses of argument*. Cambridge, England: Cambridge university press.
Trowler, P. (2011). *Researching your own institution: Higher education*.
Tursucu, S. (2019). *Successful transfer of algebraic skills from mathematics into physics in senior pre-university education.* (PhD.). University of Technology, Delft, Delft. Retrieved from https://repository.tudelft.nl/islandora/object/uuid:80f98acd-dc72-4aa8-bec6-ce72a26c2c65
United Kingdom Department for Education. (2014). National curriculum in England: Science programmes of study. In: Crown Publishing London.
van Buuren, O. (2014). *Development of a modelling learning path.* (PhD.). Universiteit van Amsterdam, Amsterdam. Retrieved from https://dare.uva.nl/search?identifier=a56fa0b2-7b6e-4b45-a712-721d9669f959
Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006). *Educational design research*. Abingdon, Oxon: Routledge.
van den Berg, E. (2013). The PCK of Laboratory Teaching: Turning Manipulation of Equipment into Manipulation of Ideas. *Scientia in educatione, 4*(2), 74-92. Retrieved from https://ojs.cuni.cz/scied/article/download/86/72/0
van den Berg, E., Buning, J., & Smits, T. (1996). Leren onderzoeken in het voortgezet onderwijs. *Nederlands Tijdschrift voor Natuurkunde*(12), 271-274.
Van Heuvelen, A. (1991). Learning to think like a physicist: A review of research-based instructional strategies. *American Journal of Physics, 59*(10), 891-897.
Van Rens, E., & Dekkers, P. (2000). Leren onderzoeken—de rol van de docent [Learning to do research—the role of the teacher]. *Tijdschrift voor Didactiek der Beta-wetenschappen, 17*(1), 76-94.
Vanderlinde, R., & Braak, J. (2010). The gap between educational research and practice: views of teachers, school leaders, intermediaries and researchers. *British Educational Research Journal, 36*(2), 299-316.
VO-raad. (2018). Scholen op de kaart (mapping schools). Retrieved from https://www.scholenopdekaart.nl/
von Kotzebue, L., Gerstl, M., & Nerdel, C. (2015). Common Mistakes in the Construction of Diagrams in Biological Contexts. *Research in Science Education, 45*(2), 193-213.
Walker, J. P., Sampson, V., & Zimmerman, C. O. (2011). Argument-driven inquiry: An introduction to a new instructional model for use in undergraduate chemistry labs. *Journal of Chemical Education, 88*(8), 1048-1056.
Walsh, C., Quinn, K. N., Wieman, C., & Holmes, N. (2019). Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. *Physical Review Physics Education Research, 15*(1), 010135.
Watson, R., Goldsworthy, A., & Wood-Robinson, V. (1999). What is not fair with investigations? *School science review, 80*(292), 101-106.
Watson, R., Swain, J., & McRobbie, C. (1999). The interaction between teaching styles and pupil autonomy in practical science investigations–a case study. *Practical Work in Science Education–Recent Research Studies*, 148-159.
Watson, R., Swain, J. R., & McRobbie, C. (2004). Students' discussions in practical scientific inquiries. *International Journal of Science Education, 26*(1), 25-45. doi:10.1080/0950069032000072764

Webb, M. (Producer). (2012, 2020/oct/08). The Amazing Spider-Man - Crane Swinging Scene. [Movie scene] Retrieved from https://www.youtube.com/watch?v=CoSY8jeLIfw
Welford, G., Harlen, W., & Schofield, B. (1985). *Assessment of performance unit: practical testing at ages 11, 13 and 15*. Retrieved from London:
Wellington, J. (2002). *Practical work in school science: which way now?* : Routledge.
White, R., & Gunstone, R. (1992). *Probing understanding*: Routledge.
Wieman, C. (2015). Comparative cognitive task analyses of experimental science and instructional laboratory courses. *The Physics Teacher, 53*(6), 349-351. doi:10.1119/1.4928349
Wieman, C. (2016). *Introductory labs; what they don't, should, and can teach (and why).* Paper presented at the APS April Meeting Abstracts, Baltimore, Maryland.
Wilcox, B. R., & Lewandowski, H. (2016). Open-ended versus guided laboratory activities: Impact on students' beliefs about experimental physics. *Physical Review Physics Education Research, 12*(2), 020132.
Wong, V. (2017). Variation in graphing practices between mathematics and science: implications for science teaching. *School science review, 98*(365), 109-115.
Wong, V. (2018). *The relationship between school science and mathematics education.* King's College London,
Woolgar, S., & Latour, B. (1986). Laboratory Life: The Construction of Scientific Facts. In: Princeton, NJ: Princeton University Press.
Woolnough, B. E., & Allsop, T. (1985). *Practical work in science*: Cambridge University Press.
Zion, M., & Mendelovici, R. (2012). Moving from structured to open inquiry: Challenges and limits. *Science Education International, 23*(4), 383-399. Retrieved from http://files.eric.ed.gov/fulltext/EJ1001631.pdf
Zwickl, B. M., Finkelstein, N., & Lewandowski, H. (2013). *Development and validation of the Colorado learning attitudes about science survey for experimental physics.* Paper presented at the AIP Conference Proceedings.
Zwickl, B. M., Hirokawa, T., Finkelstein, N., & Lewandowski, H. (2014). Epistemology and expectations survey about experimental physics: Development and initial results. *Physical Review Special Topics-Physics Education Research, 10*(1), 010120.
Zwickl, B. M., Hu, D., Finkelstein, N., & Lewandowski, H. (2015). Model-based reasoning in the physics laboratory: Framework and initial results. *Physical Review Special Topics-Physics Education Research, 11*(2), 020113.

Development of a teaching-learning sequence for scientific inquiry through argumentation in secondary physics education

Enabling students to engage in independent scientific inquiry is a highly valued but seemingly elusive goal of (secondary school) science education. Therefore, this study aims to determine and understand how to effectively develop inquiry knowledge in students. The chosen approach to enable students to plan, carry out and evaluate a physics inquiry, is to regard an inquiry as the construction of a scientifically cogent argument for a specific claim. In an authentic scientific inquiry, the researcher invests - from the very start of the inquiry - time and effort in making the inquiry's claim as indisputable as possible. The researcher strives for optimal cogency of the argument in support of that claim. Throughout the various studies in this thesis it is argued that this idea can be translated to classroom situations: fostering the insight that students' inquiry should result in a complete, correct and substantiated answer to the research question. It is shown that this is a meaningful strategy in enabling them to engage in independent scientific inquiry: it results in a cognitive need in students to develop the knowledge that allows them to produce such an answer. As such, this thesis shows that argumentation is an indispensable part of teaching scientific inquiry. Explicit attention for argumentation promotes development of students' inquiry knowledge.